STATISTICS CANADA

NATIONAL POPULATION HEALTH SURVEY

HOUSEHOLD COMPONENT

CYCLES 1 to 8 (1994/1995 to 2008/2009)

LONGITUDINAL DOCUMENTATION

July 2010

Statistics Canada Statistique Canada

Canada

Note to Users

This documentation accompanies the release of eight cycles of longitudinal data for the Household Component of the National Population Health Survey (NPHS).

This document provides a wide range of information on the NPHS: objectives, content, sample design, collection, processing, weighting procedures, data quality, tabulation guidelines and data access. Chapters 7, 8, and 11 give more details on the various subsets of respondents and their associated sampling and bootstrap weights.

This document sometimes refers to a specific cycle of the NPHS by using the years in which it occurred. Here are the NPHS cycles with their corresponding years:

Cycle 1 = 1994/1995
Cycle 2 = 1996/1997
Cycle 3 = 1998/1999
Cycle 4 = 2000/2001
Cycle 5 = 2002/2003
Cycle 6 = 2004/2005
Cycle 7 = 2006/2007
Cycle 8 = 2008/2009

This document is also intended for the share file users, i.e., the provincial ministries of health, Health Canada and the Public Health Agency of Canada. The share file includes two subsets (square and full—see Section 7), which include respondents who agreed to share the information collected as part of the NPHS (all cycles) and the corresponding sampling and bootstrap weights. These subsets of respondents are also part of the master file. However, they are to be used by the share partners. The share- file users should disregard references specific to other subsets of respondents.

Quick Links to Statistics Canada Web Pages
NPHS, Household Component—Longitudinal

Documentation, NPHS Household Component—Longitudinal

- Cycle 8—Definitions, methods and documentation

Questionnaires

- NPHS Questionnaire, Household Component, Longitudinal, Cycle 1 (1994/1995)
  http://www.statcan.gc.ca/imdb-bmdi/instrument/3225_Q1_V1-eng.pdf
- NPHS Questionnaire, Household Component, Longitudinal, Cycle 2 (1996/1997)
  http://www.statcan.gc.ca/imdb-bmdi/instrument/3225_Q1_V2-eng.pdf
- NPHS Questionnaire, Household Component, Longitudinal, Cycle 3 (1998/1999)
  http://www.statcan.gc.ca/imdb-bmdi/instrument/3225_Q1_V3-eng.pdf
- NPHS Questionnaire, Household Component, Longitudinal, Cycle 4 (2000/2001)
  http://www.statcan.gc.ca/imdb-bmdi/instrument/3225_Q1_V4-eng.pdf
- NPHS Questionnaire, Household Component, Longitudinal, Cycle 5 (2002/2003)
  http://www.statcan.gc.ca/imdb-bmdi/instrument/3225_Q1_V5-eng.pdf
- NPHS Questionnaire, Household Component, Longitudinal, Cycle 6 (2004/2005)
  http://www.statcan.gc.ca/imdb-bmdi/instrument/3225_Q1_V6-eng.pdf
- NPHS Questionnaire, Household Component, Longitudinal, Cycle 7 (2006/2007)
  http://www.statcan.gc.ca/imdb-bmdi/instrument/3225_Q1_V7-eng.pdf
- NPHS Questionnaire, Household Component, Longitudinal, Cycle 8 (2008/2009)
  http://www.statcan.gc.ca/cgi-bin/imdb/p2SV.pl?Function=getInstrumentLink&SurvId=3225&SurvVer=0&InstaId=15280&InstaVer=8&lang=en&db=IMDB&adm=8&dis=2

NPHS Internet Publication

- Healthy Today, Healthy Tomorrow? Findings from the National Population Health Survey

Contents

Page

1. Introduction 1
2. Background 2
3. Objectives 3
4. Survey Content 4
   - 4.1 Content Selection Criteria 4
   - 4.2 Cycle 8 (2008/2009) Changes to Existing Content 5
   - 4.3 Previous Interview Variables Used as Additional Information in Cycle 8 5
5. Sample Design 6
   - 5.1 Cycle 1 (1994/1995) Sample Design 6
     - 5.1.1 Sample Allocation 6
     - 5.1.2 The Rejective Method 7
     - 5.1.3 Sample Selection 7
     - 5.1.4 Sample Design in Quebec 8
   - 5.2 Longitudinal Sample 9
6. Data Collection 11
   - 6.1 Questionnaire Design and Data Collection Method 11
   - 6.2 Tests 11
   - 6.3 Interviewing 11
   - 6.4 Non-response and Tracing 12
7. Data Processing 14
   - 7.1 Editing 14
   - 7.2 Coding 14
   - 7.3 Derived and Grouped Variables 15
   - 7.4 Estimation and Weighting 16
   - 7.5 Definition of the Longitudinal Response Pattern 16
   - 7.6 Definition of Full/Complete Response and Non-response 17
   - 7.7 Subsets of Respondents 17
8. Weighting 18
   - 8.1 Longitudinal Weighting 18
     - 8.1.1 Starting Point: Cycle 1 (1994/1995) Stripped Weights 19
     - 8.1.2 Adjustments to Create the Different Longitudinal Weights 19
       - 8.1.2.1 Longitudinal Square Weight (WT64LS) 19
       - 8.1.2.2 Longitudinal Full Weight (WT6CLF) 19
       - 8.1.2.3 Longitudinal Square Share Weight (WT6CSLS) 28
       - 8.1.2.4 Longitudinal Full Share Weight (WT6CSLF) 30
9. Data Quality 31
   - 9.1 Sampling Errors 31
     - 9.1.1 Bootstrap Method for Variance Estimation 31
   - 9.2 Non-sampling Errors 32
     - 9.2.1 Response Rates 33
       - 9.2.1.1 Cycle 1 (1994/1995) Response Rates 33
       - 9.2.1.2 Cycle 2 (1996/1997) and Cycle 3 (1998/1999) Response Rates 34
       - 9.2.1.3 Cycle 4 (2000/2001), Cycle 5 (2002/2003), Cycle 6 (2004/2005), Cycle 7 (2006/2007) and Cycle 8 Response Rates 36
     - 9.2.2 Refusal Rates 38

9.2.3 Unable-to-trace Rates 39
9.2.4 Attrition Rates 39
9.2.4.1 Attrition Rates Based on the Longitudinal Full Subset 40
9.2.5 Item Refusal and "Don't Know" Rates 42
9.2.5.1 Refusal and "Don't Know" Rates by Item 42
9.2.5.2 Refusal and "Don't Know" Rates by Respondent 44
10. Guidelines for Tabulation, Analysis and Release 46
10.1 Rounding Guidelines 46
10.2 Sample Weighting Guidelines for Tabulation 46
10.2.1 Definitions of Types of Estimates: Categorical versus Quantitative 47
10.2.2 Tabulation of Categorical Estimates 48
10.2.3 Tabulation of Quantitative Estimates 48
10.3 Guidelines for Statistical Analysis 48
10.4 Release Guidelines 49
11. Using the Longitudinal Master File 51
11.1 Use of Longitudinal Weights 51
11.2 Ensuring the Reliability of Estimates Using Bootstrap Weights 51
11.3 Variable-naming Convention 52
11.3.1 Variable Name Component Structure 52
11.3.2 Positions 1-2: Variable Name / Questionnaire Section Name 53
11.3.3 Position 3: Survey Type 55
11.3.4 Position 4: Cycle (years) 55
11.3.5 Position 5: Variable Type 55
11.3.6 Positions 6-8: Variable Name 56
12. Access to NPHS Data 57
12.1 Research Data Centres 57
12.2 Remote Access 57
12.3 Data Liberation Initiative 57
12.4 Analytical Reports and Tabulations 57
13. An Analytical Technique for Longitudinal Survey Data 59
13.1 Cycle Twinning Approach 59
13.2 Creation of the Modified Subset 62
13.3 Methodological Aspects of the Cycle Twinning Approach 63
13.4 An Example of How to Use the Cycle Twinning Approach: Quitting Smoking 64

Tables

Table 1	Longitudinal Sample Size by Province
Table 2.A	Distribution of Deaths by Year of Death
Table 2.B	Subsets of Respondents
Table 3.A	Subsets of Respondents, Corresponding Sampling Weights, and Flags
Table 3.B	Variables for Cycle 2 Non-response Adjustment
Table 3.C	Variables for Cycle 3 Non-response Adjustment
Table 3.D	Variables for Cycle 4 Non-response Adjustment
Table 3.E	Variables for Cycle 5 Non-response Adjustment
Table 3.F	Variables for Cycle 6 Non-response Adjustment
Table 3.G	Variables for Cycle 7 Non-response Adjustment
Table 3.H	Variables for Cycle 8 Non-response Adjustment
Table 3.I	Variables for the Non-response Adjustment
Table 4.A	Relevant Information for Calculation of Response Rates for Cycle 1
Table 4.B	Cycle 1 Response Rates
Table 4.C	Relevant Information for Calculation of Response Rates for Cycles 2 and 3
Table 4.D	Panel Response Rate for Cycles 2 and 3
Table 4.E	Relevant Information for Calculation of Response Rates for Cycles 4 to 8
Table 4.F	Panel Response Rate for Cycles 4, 5, 6, 7 and 8
Table 4.G	Refusal Rates by Cycle
Table 4.H	Unable to Trace Rates by Cycle
Table 4.I	Attrition Type by Cycle—Longitudinal Full Subset of Respondents
Table 4.J	Refusal and “Don’t Know” Rates by Module
Table 4.K	Refusal and “Don’t Know” Rates by Respondent
Table 5	Sampling Variability Guidelines
Table 6.A	Subsets of Respondents and Corresponding Bootstrap Weights Files
Table 6.B	“Constant” Longitudinal Variables
Table 7.A	Example of Profiles of Response from Fictitious NPHS Respondents
Table 7.B	Attrition Rates
Table 7.C	Example of the Modified Subset Structure for Twinning

Appendices

Appendix A:	NPHS Household Component, Changes to the Questionnaire for Cycle 8 (2008/2009)
Appendix B:	NPHS Household Component, Examples of Cycle 8 (2008/2009) Data Feedback and Follow-up Questions

Other Reference Documents

Questionnaire
Record Layout
Alphabetical Index, Topical Index
Data Dictionary
NPHS, Derived Variables Documentation, Cycles 1 to 8
NPHS, Derived Variables List, Cycles 1 to 8
NPHS, Content Summary, Cycles 1 to 8

1. Introduction

The National Population Health Survey (NPHS) is designed to collect longitudinal information on the health of the Canadian population and related socio-demographic information. The first cycle of data collection took place in 1994/1995. The survey will continue every second year thereafter for 10 cycles. The NPHS fulfilled both cross-sectional and longitudinal needs during its first three cycles; with Cycle 4 (2000/2001) the NPHS Household component became strictly a longitudinal survey. The cross-sectional component of the Population Health Surveys Program has been taken over by the Canadian Community Health Survey (CCHS).

The NPHS is now composed of only one component: the Household component. Since 2000/2001, the North component has been conducted by CCHS rather than NPHS. After five collection cycles (1994/1995 to 2002/2003), the Health institutions component was terminated due to the large number of deaths in the sample.

The target population of the NPHS Household component includes household residents in the ten Canadian provinces in 1994/1995 excluding persons living on Indian reserves and Crown lands, residents of health institutions, full-time members of the Canadian Forces living on Canadian Forces bases, and residents of some remote areas in Ontario and Quebec.

The Household component of NPHS has completed eight cycles: Cycle 1 (1994/1995), Cycle 2 (1996/1997), Cycle 3 (1998/1999), Cycle 4 (2000/2001), Cycle 5 (2002/2003), Cycle 6 (2004/2005), Cycle 7 (2006/2007), and Cycle 8 (2008/2009).

The Cycle 8 NPHS Household component collected in-depth information on the health of the longitudinal respondents who were randomly selected in Cycle 1 and demographic information about all members of the longitudinal respondents' household. The questionnaire includes questions related to health status, use of health services, determinants of health, chronic conditions and activity restrictions. Socio-demographic information is also collected, such as age, sex, education, household income and labour force status.

This document has been produced to facilitate the use of Cycle 8 Longitudinal Master and Share Files from the Household component. **These files include data collected from respondents from Cycles 1 to 8.** The files are described in more detail in the following chapters.

Information requests, questions about data access (Research Data Centres, Remote Access), custom tabulations, general data support or questions about the data sets or their use should be directed to

Data Access and Information Services, Health Statistics Division
Tel: 1-613-951-1746, Email: nphs-ensp@statcan.gc.ca, Fax: 1-613-951-0792

2. Background

In the fall of 1991, the National Health Information Council recommended that an ongoing national survey of population health be conducted. This recommendation was based on consideration of the economic and fiscal pressures on the health care system and the commensurate need for information to improve the health status of Canada's population. Existing sources of health data were unable to provide a complete picture of the health status of the population and the myriad factors that affect health.

In April 1992, Statistics Canada received funding to develop the National Population Health Survey (NPHS). The survey was designed to be flexible and to produce valid, reliable and timely data. In addition, it had to be responsive to changing requirements, interests, and policies.

Since beginning in 1994, the National Population Health Survey has been providing unique information on the health of Canadians by responding to the need for information on health dynamics. The NPHS is a longitudinal survey with a sample of 17,276 individuals spread across Canada's 10 provinces. Every two years, these same individuals provide current and detailed information on their physical and mental health status, use of health care services, physical activities, workplace life and social environment. Over the years of follow-up, the data have shown how a wide range of factors can contribute to the improvement or deterioration of health.

Whereas data collected from people at a single point in time provides a snapshot, NPHS longitudinal data reveals the transitions towards good or poor health. NPHS data is rich in that it also enables evaluation of the relationships between socio-economic and demographic characteristics of individuals with their health status, and its evolution over time.

3. Objectives

The objectives of the National Population Health Survey are to

- aid in the development of public policy by providing measures of the population's health status, analysis of these measurements, and trends observed
- provide data for analytical studies that will assist in understanding the determinants of health
- collect data on the economic, social, demographic, occupational and environmental correlates of health
- enable better understanding of the relationship between health status and health care utilization, including alternative as well as traditional services
- provide information on a panel of people who will be followed over time to reflect the dynamic process of health and illness
- provide the provinces and territories and other clients with a health survey capacity that will permit supplementation of content or sample[1]
- allow the possibility of linking survey data to routinely collected administrative data, such as vital statistics, environmental measures, community variables, and health services utilization.

1. Owing to the longitudinal nature of NPHS, the sample option is no longer available. CCHS is now providing this possibility.

4. Survey Content

The National Population Health Survey's objectives, described in Chapter 3, provide a broad direction, particularly concerning the type of information to be collected. The first section of this chapter discusses the general criteria used for selecting survey content, and summarizes the questionnaire sections. The next section of this chapter summarizes the changes made to the NPHS content for Cycle 8. The last section presents the variables from previous interviews that are used as additional information in collection for Cycle 8.

The content of the NPHS has been quite stable since Cycle 6. Considering the longitudinal nature of the survey, no new content was added since then. However, some previous modules or questions have been repeated in the questionnaire from cycle to cycle to enable users to do comparisons over time.

NOTE: On page iii, see the links to the NPHS questionnaires and to NPHS documentation, including a content summary for the eight cycles, available on Statistics Canada's website.

4.1 Content Selection Criteria

The NPHS content was selected according to the following criteria:

- Information should relate to and help monitor the health objectives of the provinces. Where health objectives have not been established, for example at the national level, policies and programs could be considered in the selection of survey content.
- The information should not duplicate data available from other sources.
- To better understand health and its determinants, information collected should provide new knowledge in areas that have not been adequately studied.[2]
- The survey should focus on behaviours or conditions amenable to prevention, treatment, or intervention.
- The survey should collect information about conditions that impose the greatest burden, in terms of suffering or cost, on affected individuals, the general population, or the health care system.
- The survey should collect information on factors related to good health, not just those related to illness.

In Cycle 1, one person in each household was randomly selected as the longitudinal respondent to answer an in-depth questionnaire on health (Health component—H06). During the first three cycles, some information (demographic and health) was also collected about all member of the longitudinal respondent's household (General component—H05) and about the dwelling.

Starting in Cycle 4, the General and Health component questionnaires were combined into a unique questionnaire completed by the longitudinal respondents. Only demographic

2. No new content was added to NPHS questionnaire since Cycle 6. This criterion is no longer valid.

information on age, sex, marital status and relationships is collected about the other members of the household. Information about the dwelling is also collected.

Reflecting the above criteria, the NPHS questionnaire includes questions related to health status, use of health services, determinants of health, chronic conditions and activity restrictions, as well as demographic and socio-economic status. For example, health status is measured through questions on self-perception of health, functional ability, chronic conditions, and activity restriction. The use of health services is measured through questions on visits to health care providers (traditional and alternative), hospital care and on use of drugs and other medications. Health determinants that are explored include smoking, alcohol use and physical activity. Questions are asked on preventive tests and examinations, which probed for frequency and reasons for use. Demographic and socio-economic information include age, sex, education, ethnicity, race, household income and labour force status.

4.2 Cycle 8 (2008/2009) Changes to Existing Content

Questions on health are asked first, followed by the socio-economic questions (language, education, labour force status, and income), as was done in the previous cycle.

The Cycle 8 focus content was incorporated in the most suitable place in the questionnaire. The Cycle 8 focus content includes self-esteem, which was part of Cycle 4, and sense of coherence, which was part of Cycle 3.

Appendix A details changes made to the Cycle 8 questionnaire.

4.3 Previous Interview Variables Used as Additional Information in Cycle 8

To reduce respondent burden, questions to which the answer was already known and that would not change over time (e.g., country of birth) are not repeated.

Variables that could change over time if certain actions had occurred (e.g., level of education), were updated only if appropriate.

Some answers from earlier cycles were repeated to respondents in the Cycle 8 interview. This proved to be a valuable tool, resulting in better quality collected data. For instance, previous information on selected chronic conditions was recalled for the respondent in order to explain any change. For more information, please see Appendix B.

5. Sample Design

The target population of the National Population Health Survey's Household component includes household residents in the 10 Canadian provinces in 1994/1995, excluding persons living on Indian Reserves and Crown Lands, residents of health institutions, full-time members of the Canadian Forces living on Canadian Forces bases, and residents of some remote areas in Ontario and Quebec. This chapter describes the Cycle 1 sample design and explains how the sample of 17,276 persons was selected.

5.1 Cycle 1 (1994/1995) Sample Design

The Labour Force Survey (LFS) sample design, redesigned in 1991, was used as the basis for the sample design in all provinces except Quebec, where the NPHS sample was selected from households already being interviewed by *Santé Québec* for the 1992–1993 *Enquête sociale et de santé* (ESS).
Three factors shaped the sample design of the household component sample:

- the targeted national and provincial sample sizes;
- the decision to select one member per household to make up the longitudinal panel;
- the choice of the LFS sample design as a vehicle for selecting the sample.

These three factors resulted, respectively, in the allocation of the sample, the application of the 'rejective method,' described later, to improve the sample's representativeness, and the selection of provincial samples outside Quebec.

5.1.1 Sample Allocation

The NPHS initially had a target sample size of 19,600 respondent households. Further, national and provincial representatives agreed that each province needed a minimum of 1,200 households. Subject to this restriction, the provincial sample sizes were obtained by using a well-known allocation scheme that balances the reliability requirements at national and regional levels (Kish, 1988).[3] According to this scheme the sample was allocated proportionally to $\sqrt{(0.804W_h^2 + 1/12^2)}$, where W_h is the 1991 Census proportion of households in province h, h=1,...,10. This allocation determined the base sample size for each province. Four provinces chose to increase their allotted sample size for the Cycle 1 through the option of buy-in of additional units with increased funding, for cross-sectional purposes. These additional units were not retained for the longitudinal sample.

3. Kish, L. (1988). Multipurpose Sample Design, *Survey Methodology*, 14, 19-32.

5.1.2 The Rejective Method

The survey content primarily focused on one member in each sample household who was chosen at random to become the longitudinal panel respondent. Without the use of the rejective method, the panel would under-represent persons coming from large households, typically parents and children, since they had less chance of being chosen and over-represent persons coming from small households, often single people or the elderly.

Thus, a rejective method was adopted to increase the representation of parents and youths in the panel. To do so, a portion of the sample was pre-identified for screening. After their member roster was completed, screened households that had no member younger than 25 years of age were eligible for rejection and dropped out of the survey. In order to maintain the required sample sizes, the number of households visited in each province was increased by the anticipated number of households screened out in this way.

The rejective method with an under-25-year-old rule was adopted, as it performed better than other rejection rules considered. For cost and operational reasons, the percentages of preliminary screened households was usually limited to 25% to 30% in Ontario, 37.5% to 40% in urban areas elsewhere and 25% to 30% in rural areas. As apartment strata had a high concentration of small households, their sample sizes were reduced instead of applying a rejective method. The rejective method was also not applied in remote regions because of the high contact costs there.

5.1.3 Sample Selection

The sample design considered for the household component of the NPHS was a stratified multi-stage design. In the first stage, homogeneous strata were formed and independent samples of clusters were drawn from each stratum. In the second stage, a dwelling list was prepared for each chosen cluster, and some were selected from the list.

In all provinces except Quebec, the NPHS used the multi-purpose sampling methodology developed for the redesign of the LFS. That methodology provided general household surveys with clustered samples of dwellings, making the sample design very cost-effective for listing and collecting data.

The basic LFS design is a multi-stage stratified sample of dwellings selected within clusters. Each province is divided into three types of areas (Major Urban Centres, Urban Towns and Rural Areas) from which separate geographic and/or socio-economic strata are formed. In most strata, six clusters, usually Census Enumeration Areas, were selected with Probability Proportional to Size. In a few cases where the population density was low, an additional stage was added by first selecting two or three large Primary Sampling Units, dividing them into clusters, and drawing a sample of six clusters from each. The number 6 was used throughout the sample design to allow a one-sixth rotation of the sample every month for the LFS.

The sample of dwellings was obtained after listing operations in sampled clusters were completed. As sampling rates were predetermined, there were often differences between anticipated and obtained sample counts. Excessive sample yields were corrected by dropping a portion of the originally selected units. This was usually done at aggregated levels, and was called sample stabilization. Note also that sample sizes were inflated to represent dwellings rather than households, as a certain amount of non-response was expected, and a portion of the dwellings were expected to be vacant or otherwise out of scope.

The LFS sample design is set up to yield about 60,000 households. Surveys needing smaller sample sizes usually 'reserve' from one to six rotations per province, a rotation being one-sixth of the total sample. Sample stabilization is used to maintain the sample at a desired level, as when two rotations are reserved but the sample size needed only represents 1.5 rotations.

Requirements specific to the NPHS led to two modifications to this sampling strategy. The number of reserves needed was specified at the stratum level rather than the provincial level. This was done in order to meet the specific sub-provincial sample size requirements for cross-sectional purposes in the first cycle. Also, the number of clusters selected per stratum had to be a multiple of four for purposes of variance estimation and seasonal representativeness (this allowed strata to have two or more independent samples of four clusters each—one per collection period). As the NPHS usually requested only between two and six clusters per LFS stratum, similar LFS strata were grouped to form larger NPHS strata with the required number of sample clusters. Once strata were grouped, their sample clusters were also grouped to form replicates.

As a result of these modifications, the NPHS sample of clusters can be considered as a stratified replicated sample, where strata are groups of LFS strata and replicates are typically independent, identically distributed samples of four clusters each. There were exceptions, but they are not expected to have a significant impact on survey results. Two design variables named "Stratum" and "Replicat" can be found on the Master file, where Stratum represents the LFS stratum, and Replicat represents the NPHS replicates.

5.1.4 Sample Design in Quebec

In Quebec, the NPHS sample was selected from dwellings participating in a *Santé Québec* health survey: the 1992/1993 ESS. The survey sampled 16,010 dwellings using a two-stage sample design similar to that of the LFS. The province was divided geographically by crossing 15 health areas with four urban density classes (the Montréal Census Metropolitan Area, regional capitals, small urban agglomerations and the rural sector). In each area, clusters were stratified by socio-economic characteristics and were selected using a PPS sample. Selected clusters were enumerated, and random samples of their dwellings were drawn: 10 per cluster in major cities, 20 or 30 elsewhere.

Santé Québec provided non-confidential information that enabled classification of their sample into four types of households: one-member households; households with children; other households with youths (persons aged under 25); and the rest

(more than one member and no youth or child). A household type was determined by NPHS personnel for the ESS non-respondents.

The NPHS sample size was first allocated among the four urban density classes. To avoid having too much sample in Montréal, the allocation was proportional to $\sqrt{(2W_h^2 + 1/4^2)}$, where W_h is the population share for class *h, h*=1,2,3,4. In each class, an attempt was made to obtain a sub-sample from the ESS that, as far as the selected panel member was concerned, would be proportional to the populations for the four household types. This was done by drawing a sufficient number of households from the ESS to give the required yield for households with children—the most under-represented group—and then removing excess sample from the other three household groups. An initial sample, which was almost 50% higher than needed, was thus selected. After removing from it 2/3 of the one-member households, 1/2 of the other households with no youths or children, and 1/6 of households with youths but no children, the objective was nearly attained.

Considerations for seasonal representation and variance estimation, and integration with the National Longitudinal Survey of Children and Youth (NLSCY), affected the sub-sampling in Quebec as they did elsewhere. ESS strata were thus collapsed to allow the formation of replicates, with the clusters in each replicate covering all four quarters (two quarters are covered per cluster in the rural and small urban sectors, as sample sizes are higher there). The sample of households with children was split into an "adult" sample and a "children" sample by a 3:2 ratio, the terms having the same meaning as in other provinces. "Children" sample households in quarters 1 and 2 were reassigned to quarters 3 and 4. Since NPHS surveyed the current occupants of dwellings selected for the ESS, and changes occurred in some of those dwellings, the samples of households without children for quarters 3 and 4 were also to be split, by a 2:3 ratio, into an "adult" and a "children" sample.

5.2 Longitudinal Sample

The longitudinal sample, also called the longitudinal panel or simply the panel, is composed of the 17,276 persons that were selected in Cycle 1 and have completed at least the General component of the questionnaire in Cycle 1. It also includes 2,022 children from the first cycle (1994/1995) of the National Longitudinal Survey of Children and Youth (NLSCY). These children were interviewed by the NLSCY for the NPHS in Cycle 1, and have been interviewed by the NPHS since the second cycle. This longitudinal sample (17,276), has been and will be surveyed in all NPHS cycles. Additional samples added to Cycles 1, 2 and 3 for cross-sectional purposes are not part of the longitudinal sample.

The longitudinal sample is not renewed over time. No panel members were or are to be classified out-of-scope. The longitudinal sample size remains the same (17,276) for all cycles. Consequently, for Cycle 8 all longitudinal panel members were 13 years old and older, and the longitudinal sample did not contain anyone who has immigrated to Canada after 1994/1995.

The number of people answering the survey decreases slightly from one cycle to the next, due to attrition caused by non-response (for example, refusals and individuals that were untraceable). Despite the attrition, the longitudinal sample is still representative of the

1994/1995 population. This relatively small attrition (see Section 9.2.4) should not lead to large increases in the variance of estimates. Note that panel members who died and panel members who moved to a health institution are still part of the longitudinal sample, and are considered as respondents (see Section 7.6). Therefore, these persons do not contribute to the attrition of the NPHS longitudinal panel.

Table 1 presents the sample size of the longitudinal sample by province in 1994/1995. It also shows the number of people that provided a full response to all eight NPHS cycles.

Table 1: Longitudinal Sample Size by Province

Province	Longitudinal sample Cycle 1 (1994/1995)	Number of respondents full response in Cycles 1 to 8
Newfoundland and Labrador*	1,082	657
Prince Edward Island	1,037	643
Nova Scotia	1,085	647
New Brunswick	1,125	633
Quebec	3,000	1,766
Ontario	4,307	2,323
Manitoba	1,205	728
Saskatchewan	1,168	747
Alberta	1,544	886
British Columbia	1,723	952
Total	**17,276**	**9,982**

* Labrador was not part of Newfoundland when sample was selected.

6. Data Collection

6.1 Questionnaire Design and Data Collection Method

The survey questions were designed for computer-assisted interviewing (CAI). Thus, as the questions were developed, the associated logical flow into and out of the questions was specified, along with the type of answer required, the minimum and maximum values, online edits associated with the question, and what to do in case of item non-response.

With CAI, the interview is controlled based on answers provided by the respondent. On-screen prompts are shown when an invalid entry is recorded: this gives immediate feedback to the respondent or the interviewer to correct inconsistencies. Another advantage of CAI is automatic insertion of reference periods based on current dates. Pre-filling of text or data based on information gathered during the current interview or previous cycles' interviews enables the interviewer to proceed without having to search back for previous answers. This type of pre-fill includes such things as using the correct name or sex within the questions themselves. Allowable ranges/answers based on data collected during the interview can also be programmed. In other words, the questionnaire is customised to the respondent based on the data collected in the current interview and in previous cycles.

6.2 Tests

The CAI application was extensively tested in-house in order to identify any errors in the program flow and text. For Cycles 1 to 6, two field tests were conducted. The tests involved four of Statistics Canada's regional offices. The main objectives of the two tests were to observe respondent reaction to new questions and new modules, estimate the time needed for the various sections, study response rates and test feedback questions. Field operations and procedures, interviewer training, and the CAI application (i.e., the questionnaire on computer) were also tested. Since Cycle 7, only one field test was conducted (the samples from the two field tests were combined). The majority of long-term non-respondents were removed from the sample. A small number was kept in order to be able to test tracing procedures among other things. The objectives of the test remained the same as before.

6.3 Interviewing

In Cycle 8, data collection for the household component was divided into four quarters (starting in June and continuing in July and September 2008, and January 2009). An additional collection period was held starting in May 2009 to follow up with non-respondents from previous quarters.

In Cycle 8, interviewers working in Statistics Canada calling centres located in Edmonton, Sturgeon Falls, Sherbrooke, and Halifax performed data collection.

A special collection is conducted for the panel members residing in health care institutions. The interview for these respondents was conducted in person. A paper questionnaire was used for the interviews in Cycles 1 to 7. In Cycle 8, a computer-assisted personal interview was conducted: 267 respondents residing in health care institutions were interviewed. They are identified as such on the master and share files (see Section 7.5). However, the health

institutions component questionnaire collects less information than the household component questionnaire. Missing variables for these respondents are coded to "6", i.e., not applicable, on the household component files.

All interviewers were employees hired and trained by Statistics Canada specifically to carry out surveys. National Population Health Survey (NPHS) data collection was performed under the supervisory and control structure put in place by Statistics Canada. All interviewers attended a training session that focused on NPHS content, and they received an Interviewer's manual for use as a reference tool.

Each living longitudinal panel member received by mail a letter announcing the start of NPHS Cycle 8 data collection. At the same time, the household component's respondents received a brochure, NPHS News (a small newspaper-style brochure with highlights of survey results) and a bookmark/ruler/magnifier to thank them for their participation. Furthermore, NPHS information for survey participants was also available on the Statistics Canada website.

In Cycles 6, and 7, a follow-up survey was conducted with a sub-sample of the November field test respondents. The objective of this follow-up was to verify with respondents

- the clearness and usefulness of the brochure, Breaking News, Body Mass Index tool, and Statistics Canada's website for survey participants
- the impact of these documents on their motivation to participate
- their interest in receiving additional NPHS information.

These follow-up surveys showed that most respondents who received the documents were satisfied and read them. However, the impact on their participation is minimal. The BMI tool was modified based on received suggestions. The majority of consulted respondents wanted to receive additional survey information. Starting in Cycle 6, a question was added at the end of the questionnaire to identify these respondents. When the thank-you material is sent, additional information is included for interested respondents.

Respondents from the household component are generally contacted by telephone. In fact, 99% of the interviews in Cycle 8 were done over the telephone. Personal visits were made if the respondent did not have a telephone, if the interviewer made a personal visit in the course of tracing a respondent, upon request by the respondent, or if the respondent resided in a health care institution. The total interview time averaged just under an hour.

Information about all household members (age, sex, and relationships between members) was obtained from the longitudinal respondent. Proxy reporting for longitudinal respondents aged 12 and older was allowed only for reasons of illness or incapacity. Such proxy reporting accounted for 5.7% of the information collected for respondents aged 12 and older. However, all interviews for respondents under 12 years old were done by proxy.

6.4 Non-response and Tracing

Many strategies were put in place to reduce the number of non-response cases. For example, the maximum assignment size for an interviewer was set, to avoid overburdening interviewers: it was based on experience from previous cycles. This allowed for the efficient follow-up of non-contact cases. Interviewer training covered ways of reducing the number of non-contacts (e.g., telephoning or visiting at various times of the day) using contact information given in the previous interview.

Interviewers were instructed to make all reasonable attempts to obtain NPHS interviews with longitudinal respondents. For cases in which the timing of the interviewer's call (or visit) was inconvenient, an appointment was made to call back (or come back) at a more convenient time. If no one was home, numerous callbacks were made. For individuals who refused to participate in the NPHS, a letter was sent from the regional office to the respondent, stressing the importance of the survey and the respondent's co-operation. This was followed by a second call (or visit) from the interviewer.

Refusals were followed up by senior interviewers, project supervisors or other interviewers to try to convince respondents to participate in the survey. To maximize the response rate, many non-response cases were also followed up in subsequent collection periods.

The failure to trace a longitudinal respondent was another type of non-response. Interviewers used several methods to trace a respondent. The last known address and telephone number were provided as part of the information on the case, as well as the name and address of one or two other contacts, if collected in a previous cycle. In addition, interviewers were trained to follow up available public leads such as local telephone directories and directory assistance. If these leads were unsuccessful, the case was transmitted to an interviewer specially trained in tracing respondents. Tracer interviewers had access to Canada-wide telephone directories and reverse directories. The cumulative non-response rate due to failure to trace the longitudinal respondent is 7.5% of the total panel, which is relatively low for the eighth cycle of a survey. Section 9.2.3 presents non-response rates due to non-tracing with more details.

Attempts were made to contact panel members who had moved within Canada or to the United States. For panel members living outside Canada and the United States, attempts were made to confirm their place of residence. The survey was not conducted if these members were still living outside Canada and outside the United States; however, information was updated for the next cycle.

7. Data Processing

7.1 Editing

Editing was first performed online in the Computer Assisted Interview (CAI) application during data collection. It was not possible to enter out-of-range values, and flow errors were controlled by the skip pattern programmed in the CAI system. For example, CAI ensured that questions that did not apply to the respondent were not asked. In the case of contradictory responses between questions, warning messages were invoked. In some situations, the conflict had to be resolved before the interview could continue. In other situations, the contradiction was accepted and no corrective action had to be taken. Because of such cases, edits were developed to be performed after data collection at head office. Inconsistencies were usually corrected by setting one or more variables to "not stated." No imputation was done.

7.2 Coding

For several National Population Health Survey (NPHS) questions, a long answer was captured (e.g., medication, illness, occupation, place of work). During data processing, the responses to these questions were coded according to the standard classification systems used by Statistics Canada.

After the dissemination of microdata from Cycles 4 through 7, the industry and occupation data for all cycles were coded to the North American Industrial Classification System 1997, and the Standard Occupational Classification 1991. For the Cycle 8 (2008/2009) microdata file, industry and occupation data for all cycles were coded to the North American Industry Classification System 2007, and the National Occupational Classification—Statistics (NOC-S) 2006.

Since the release of the Cycle 5 microdata file, the drug coding for all cycles is based on the Anatomical Therapeutic Chemical classification developed by the World Health Organization, as available on the Health Canada Drug Product Database (September 2005). A complete list of the codes is available on request.

Starting with the release of the Cycle 6 microdata file, the conditions or health problems causing activity restrictions for all cycles are coded based on the International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10). Furthermore, the Musculoskeletal Impairment Supplementary Coding Scheme is no longer used, and it is not possible to code using ICD-9 at Statistics Canada.

Many NPHS questions offer response categories, and also permit capture of a long answer. During data processing, when appropriate, these long answers are coded to existing categories. Occasionally, mainly in Cycle 1, new categories were created.

The death of a longitudinal panel member is verified against the Canadian Vital Statistics Database—Deaths when possible. Once the death is confirmed, the cause of death and date of death, if different from the one collected during the survey, are captured. The cause of death code that is captured from the Canadian Vital Statistics Database is an ICD-10 code. Variables (other than date of death and cause of death) for deceased panel members are set to “9” (i.e., not stated) in the microdata file.

A total of 2,373 panel members died during the first 8 cycles of NPHS. Among this number, 2,182 deaths have been confirmed with the Canadian Vital Statistics Database—Deaths. The year of death is known for 179 of the 191 remaining cases. Therefore, for 12 cases neither the year nor the cause of death is known. The following table presents the number of deaths and the number of confirmed death per year of death.

Table 2.A: Distribution of Deaths by Year of Death

Year of death	Number of deaths	Number of deaths confirmed with the Canadian Vital Statistics Database—Deaths
1994	25	25
1995	115	115
1996	151	148
1997	147	144
1998	188	184
1999	161	158
2000	180	176
2001	154	151
2002	182	181
2003	175	173
2004	189	187
2005	177	171
2006	203	196
2007	176	173
2008	108	0
2009	30	0
Unknown	12	0
Total	**2,373**	**2,182**

7.3 Derived and Grouped Variables

To simplify data analysis, many variables on the file have been derived using items found on the NPHS questionnaires. For example, several variables may be combined to create a new derived variable. Derived variable names generally have a *D* as the fifth character of the variable name—see Section 11.3 for more detail on the variable naming convention. In other cases, see the document, “National Population Health Survey—Derived Variables Documentation—Cycles 1 to 8” for details on how these variables were derived.

Grouped variables were created from certain variables; i.e., the values of the variable have been grouped in order to create another variable. In some cases, the derived variables are straightforward, involving collapsing response categories. Grouped variable names

generally have a *G* as the fifth character of the variable name—see Section 11.3 for more detail on the variable naming conventions.

7.4 Estimation and Weighting

The principle behind estimation in a probability sample, such as the NPHS, is that each person in the sample 'represents', besides himself or herself, several other persons not in the sample. For example, in a simple random 2% sample of the population, each person in the sample represents 50 persons in the population. In the terminology used here, it can be said that each person has a weight of 50.

The weighting phase is a step that calculates each person's associated weight. This weight must be used to derive meaningful estimates from the survey. For example, if the number of individuals whose general health has deteriorated between two survey cycles is to be estimated, it is done by selecting the records referring to those individuals in the sample having that characteristic and summing the weights entered on those records.

The NPHS weighting method is presented in Chapter 8.

7.5 Definition of the Longitudinal Response Pattern (LONGPAT)

In each cycle, depending on the outcome of the interview, each member of the panel is assigned one of these statuses:
Completed (1): panel members who provided a complete response to the interview (i.e., answered all the questions up to a given point in the NPHS questionnaire).
Deceased (2): deceased panel members.
Institutionalized (3): members of the panel residing in health care institutions. This status indicates that a complete or partial response was obtained via the collection mechanism for the institutional component.
Partial response (4): panel members who only partly responded to the NPHS questionnaire.
Non-response (5): none of the above status was assigned.

Over the cycles, the response statuses of a given respondent are concatenated into a single variable called the 'longitudinal response pattern' (LONGPAT). This variable is available on the NPHS microdata file. It can be used to rapidly obtain the response profile of a member of the panel. This variable is also used to identify different analytical subsets, as described in Section 7.7.

During data processing of a current cycle, an error is sometimes discovered in the response status of a previous cycle. Corrections are then made to the longitudinal response pattern. For example, panel members with a response status of "non-response" the previous year are found to be deceased after their data are linked to the Canadian Vital Statistics Death Database: their response status is then set to "deceased" according to the date of death found in the database. For example, the variable LONGPAT for this person in Cycle 7 was 1155555 but it becomes 11552222 in Cycle 8.

7.6 Definition of Full/Complete Response and Non-response

Since Cycle 4, NPHS has been strictly longitudinal. The definition of a response is not the same for longitudinal and cross-sectional purposes. For the NPHS longitudinal panel, a Full/Complete response includes panel members with the following statuses: completed, deceased and institutionalized.

Therefore, non-response includes panel members with the following statuses: partly completed and non-response.

7.7 Subsets of Respondents

To provide greater flexibility to users, a single microdata master file has been created for NPHS Cycle 8. This file includes all 17,276 NPHS longitudinal panel members, regardless of their response patterns from Cycles 1 to 8. Within the master file, four subsets of respondents were created, along with corresponding sampling weights and the flags to make their identification easier. For more information regarding the calculation of each subset's sampling weights, see Chapter 8; for more on the use of longitudinal weights, see Section 11.1. Table 2.B provides a description of the four subsets of respondents based on type of response.

Table 2.B: Subsets of Respondents

Subset of Respondents	Type of Response	Flag	Number of Respondents
Longitudinal Square	Complete panel: all panel members regardless of their response pattern in Cycles 1 to 8.	None, all records	17,276
Longitudinal Full	All panel members with a complete response (Full) in Cycles 1 to 8.	W**F**6CLF	9,982
Longitudinal Square Share	All panel members regardless of their response pattern, and who agreed to share their data.	W**F**6CSLS	15,805
Longitudinal Full Share	All panel members with a complete response (Full) in Cycles 1 to 8, and who agreed to share their data.	W**F**6CSLF	9,627

Users of the share file, i.e., provincial health departments, Health Canada, and the Public Health Agency of Canada,[4] should note that the Longitudinal Square Share" subset of respondents, which includes the flag for the Longitudinal Full Share subset, is provided separately, with corresponding sampling weights, for both subsets. The sampling weights and the flags of the other subsets are not on the share file CD-ROM.

4. The federal government developed a new structure to adapt to new needs in health. Health Canada and the new Public Health Agency of Canada (PHAC) come under the Minister of Health. The PHAC is one of the NPHS sharing partners.

8. Weighting

This chapter describes the weighting procedures for each subset of respondents described in Section 7.7. The longitudinal weighting process is necessarily different from that of cross-sectional weighting, for several reasons. First, longitudinal weights must represent the probability of selection of the unit of analysis at the time of sample selection. Since the longitudinal sample was selected in 1994/1995, the weights must reflect the probability of selecting the individual in Cycle 1 and not in subsequent cycles. As well, the definition of a longitudinal response is different from that of a cross-sectional response. This necessitates different adjustments particular to each type of non-response. Analysts should always use the longitudinal weights made from the subsets of respondents. The longitudinal weights have been calculated specifically to represent the 1994/1995 target population. In Cycles 1, 2 and 3, both cross-sectional and longitudinal files were produced. Although panel members were part of the cross-sectional and longitudinal files, their weights were not identical for these two types of files; rather, they were adjusted to correctly represent the target population.

For Cycle 8, four sets of weights, WT64LS, WT6CLF, WT6CSLS and WT6CSLF, have been created. Table 3.A shows the subsets of respondents and the corresponding sampling weights and flags. A panel member is part of a given subset when the flag is equal to 1.

Table 3.A: Subsets of Respondents, Corresponding Sampling Weights, and Flags

Subset of Respondents	Sampling Weight	Flag
Longitudinal square	W**T**64LS	None, all records
Longitudinal full	W**T**6CLF	W**F**6CLF
Longitudinal square share	W**T**6CSLS	W**F**6CSLS
Longitudinal full share	W**T**6CSLF	W**F**6CSLF

Only the WT6CLF, WT6CSLS and WT6CSLF weights have been adjusted for non-response. However, all four weights were 'post-stratified' to the 1994/1995 population estimates based on the 1996 Census counts by age group[5] and sex within each province. Post-stratification is used to ensure that the four subsets of respondents correctly represent the 1994/1995 NPHS target population. The next section describes the NPHS longitudinal weighting method.

8.1 Longitudinal Weighting

The longitudinal weighting procedure is based on the weighting done for the Cycle 1 National Population Health Survey (NPHS) cross-sectional sample. Some weight adjustments were applied to the Cycle 1 cross-sectional weights in order to incorporate the additional sample used exclusively for cross-sectional purposes. These adjustments were removed for the longitudinal panel weight to create a 'stripped' weight. This stripped weight is the starting point for obtaining the longitudinal weight.

5. Post-stratification is done by using the updated date of birth instead of using the age variable at Cycle 1 (DHC4_AGE), which is never updated.

8.1.1 Starting Point: Cycle 1 (1994/1995) Stripped Weights

The Cycle 1 stripped weights were obtained using the Labour Force Survey basic weights as a starting point for all provinces except Quebec, where the basic weights from the *Enquête sociale et de santé* were taken as a starting point. Several adjustments were made to these weights, to take into account the nature of the NPHS and to accurately represent the true probability of selection for each panel member. All of the adjustments made in Cycle 1 are kept for the subsequent cycles: this is because the longitudinal sample always refers to the same population that is the population of 1994/1995.

A full description of the Cycle 1 weighting procedures still relevant for subsequent cycles is included in Sections 11.3 and 11.4 of the Cycle 2 Public-use Microdata File documentation.

http://www.statcan.gc.ca/imdb-bmdi/document/3225_DLI_D9_T22_V2-eng.pdf

At this point, adjustments were made to the stripped weight to obtain the various sets of longitudinal weights.

8.1.2 Adjustments to Create the Different Longitudinal Weights

8.1.2.1 Longitudinal Square Weight (WT64LS)

The longitudinal square weight **WT64LS** is to be used with the longitudinal square subset. It is calculated by post-stratifying the Cycle 1 stripped weight to the 1994/1995 population estimates based on 1996 Census counts by age group (0 to 11, 12 to 24, 25 to 44, 45 to 64, and 65 and older) and sex within each province. The post-stratification adjustment is given by

$$\frac{\textit{Population estimate in a province/age/sex category}}{\textit{Sum of stripped weights of respondent household members in a province/age/sex category}}$$

8.1.2.2 Longitudinal Full Weight (WT6CLF)

The longitudinal full subset includes only panel members who provided a full response, i.e., members who have a status of “complete,” “deceased” or “institutionalized” at each cycle. Panel members who are excluded from this subset were therefore non-respondents, i.e., they had a status of “partial response” or “non-response” at some point during the first eight cycles of the survey, and their weight must be redistributed to compensate for this non-response.[6]

The Cycle 1 stripped weight is the starting point and adjustments for non-response are made. A different non-response adjustment is made for each cycle, and these adjustments are cumulative from one cycle to another. For example, to obtain the Cycle 8 weights, the non-response adjustments for Cycles 2 to 8 are applied successively to the Cycle 1 stripped weights.

6. See Section 7.5 for the definition of longitudinal response pattern and Section 7.6 for the definition of full/complete for NPHS.

The adjustments were necessary in order to obtain the Cycle 8 Longitudinal Full Weight are described below.

Adjustment 1: Adjustment for Cycle 2 (1996/1997) Non-response

Adjusting for non-response was done using the weighting class approach. Weighting classes consist of groupings of respondents who share the same propensity to respond to the survey. Characteristics from Cycle 1, available for Cycle 2 respondents and non-respondents alike, are used to define membership in the weighting classes. Classes are formed using a clustering algorithm that arranges the sample units into a tree structure by successively splitting the data set into 'branches' based on the units' characteristics. Each split aims to divide the present units into two or more groups that have the most dissimilar observed non-response rates (and within which the non-response rates are expected to be more similar).

A different characteristic may be used to define each split. For example, units may first be divided into owner-occupied dwellings and rented dwellings. The former split may then be further split into five groups based on the level of household income; the latter may be further split based on the respondent's age. Each of the newly formed groups may be further split, based on other characteristics, and so on. The results of the final splits are the weighting classes.

The Chi-square Automatic Interaction Detection (CHAID) algorithm was used to determine the weighting classes. To produce more stable adjustments, a minimum of 30 units per weighting class was used.

Separate weighting classes were created for each province. Here, the province here refers to the province of residence at the time of the sample selection in 1994/1995. The Cycle 1 household characteristics and the personal characteristics of the longitudinal member were considered. Some characteristics related to the sampling design of the survey or to the sampling weight were also considered in an effort to incorporate the sampling design of the survey into the analysis. Personal characteristics from the Health component were not used because they were unavailable for many longitudinal members in 1994/1995.

The variables chosen by the CHAID algorithm to build weighting classes to adjust for Cycle 2 non-response are listed in Table 3.B. Two variables from the Cycle 1 sample design were used: a flag indicating the presence of members under 25 years old in the household, and a flag indicating the presence of members under 12 years old in the household. The Cycle 1 non-response flag for income and the flag indicating if the individual was under age 16 were also used. Please refer to the Data Dictionary for a complete description of the variables listed in Table 3.B.

Table 3.B: Variables for Cycle 2 Non-response Adjustment

DHC4_AGE	DHC4_MAR	GE34DURB	LFC4_1	SDC4DRAC
DHC4DECF	DHC4_OWN	HCC4DMDC	RAC4F1	SDC4GCB7
DHC4_DWE	GE34DCMA	INC4DIA5	SDC4DAIM	SEX

To adjust for longitudinal members who did not respond in Cycle 2, the following adjustment is applied to the weight of respondents:

$$\frac{\text{Sum of weights for all longitudinal members}}{\text{Sum of weights for Cycles 1 and 2 responding longitudinal members}}$$

This adjustment is performed within each weighting class.

Adjustment 2: Adjustment for Cycle 3 (1998/1999) Non-response

The 15,672[7] records with a full response after two cycles—those with a longitudinal response pattern 11, 12 or 13[8]—are taken as the starting point. A non-response adjustment is applied to the cases that have a status of "full response" after three cycles (defined as one of the following response patterns: 111, 112, 113, 122, 131, 132 or 133). All other response patterns, i.e., 114, 115, 134 and 135, are considered non-responses. Records for which the panel member was deceased in Cycle 2 (pattern 122) or institutionalized since Cycle 2 (pattern 133) are treated differently from the rest. For these records, no non-response adjustment is made. This is because their weight in Cycle 2 has been already adjusted to reflect the fact that some of the Cycle 2 non-respondents may have in fact been deceased or institutionalized.

Adjusting for non-response was done using the weighting class approach. Separate weighting classes were created for each province, i.e., the 1994/1995 province of residence. When adjusting for non-response in Cycle 3, only the Cycle 2 household characteristics, and the personal characteristics of the longitudinal member, were considered. Again, as for Cycle 2, characteristics related to the sampling design of the survey or to the sampling weight were considered in an effort to incorporate the sampling design of the survey into the analysis. However, while personal characteristics from the Health component were not used in the Cycle 2 non-response adjustment, they were used for Cycle 3, as they were available for all records that went into the Cycle 3 non-response adjustment.

The variables chosen by the CHAID algorithm to build weighting classes to adjust for Cycle 3 non-response are listed in Table 3.C. A Cycle 1 sample design variable indicating the household type (adult or child) was used, as was a Cycle 2 item non-response flag for income. Please refer to the Data Dictionary for a complete description of the variables listed in Table 3.C.

7. When the Cycle 2 data were released, there were 15,670 records in the longitudinal full subset. With the information from Cycle 8, we went back to previous cycles and could confirm, among other things, that some non-responses were in fact deaths and some deaths were a non-response. Some cases that had not been full/complete became so; others that had been full/complete were no longer so. These modifications raised the number of cases with a full/complete response after two cycles from 15,670 to 15,672.
8. See Section 7.5 for the definition of longitudinal response pattern and Section 7.6 for the definition of full/complete for NPHS.

Table 3.C: Variables for Cycle 3 Non-response Adjustment

AD_6_1	DHC6_AGE	INS6_4	SDC6DAIM	SMS6_9A
AD_6_7	DHC6_MAR	INS6_6	SDC6_4P	SMS6_13A
ALC6WKY	DV_6_65J	LFC6_41	SDC6_5A	SMS6_13C
ALC6_3	EDC6_3	MHC6DWK	SDC6_5F	SMS6_13E
AM56_SHA	ES_6_80	MHC6_1A	SDC6_6B	SMS6_16D
AM66_PXY	GE36LMOV	MHC6_1B	SDC6_7A	SMS6_18A
AM66_SHA	HCC6F1	MHC6_1F	SDC6_7B	SMS6_18D
BPC6_10	HSC6DPAD	MHC6_1L	SDC6_7D	SP36_CPA
CCC6DNUM	HWS6_5	MHC6_13	SEX	SSC6D2
CCC6_1L	INC6DIA5	PC_6_40	SHS6_4	SSC6_3
CCC6_1N	INC6_1A	RPC6_3	SMC6_2	SSS6_2
DGC6_1D	INC6_3B	RSS6_1	SMC6_5	SSS6_4

To adjust for longitudinal members who did not respond in Cycle 3, the following adjustment is applied to the weight of respondents:

$$\frac{\text{Sum of weights for Cycles 1 and 2 responding longitudinal members}}{\text{Sum of weights for Cycles 1, 2 and 3 responding longitudinal members}}$$

This adjustment is performed within each weighting class, and is calculated from records with the following longitudinal response patterns: 111 to 115, 131, 132, 134 and 135. Again, records for panel members who were deceased in Cycle 2 or institutionalized since Cycle 2 are not part of this adjustment.

Adjustment 3: Adjustment for Cycle 4 (2000/2001) Non-response

The 14,630[9] records with a full response after three cycles are taken as the starting point. Again, records for which the panel member was deceased in Cycle 2 or 3 or institutionalized since Cycle 2 or 3 are treated differently from the rest. For these records, no non-response adjustment is made since their weight in Cycle 2 or 3 has been already adjusted to reflect the fact that some of the Cycle 2 or Cycle 3 non-respondents may have in fact been deceased or institutionalized.

Adjusting for non-response was done using the weighting class approach. Separate weighting classes were created for each design province—i.e., the 1994/1995 province of residence. When adjusting for non-response in Cycle 4, only the Cycle 3 household characteristics as well as personal characteristics of the longitudinal member were considered. As for Cycle 3, characteristics related to the sampling design of the survey or to the sampling weight were considered in an effort to incorporate the sampling design of the survey into the analysis.

9. When the Cycle 3 data were released, there were 14,619 records in the longitudinal full subset. With the information from Cycle 8, we went back to previous cycles and could confirm, among other things, that some non-responses were in fact deaths and some deaths were a non-response. Some cases that had not been full/complete became so; others that had been full/complete were no longer so. These modifications raised the number of cases with a full/complete response after three cycles from 14,619 to 14,630.

Personal characteristics from the Health component were used, because they were available for all records that went into the Cycle 4 non-response adjustment.

The variables chosen by the CHAID algorithm to build the weighting classes to adjust for Cycle 4 non-response are in Table 3.D. A Cycle 3 item non-response flag for income was also used. Please refer to the Data Dictionary for a complete description of the variables listed in Table 3.D.

Table 3.D: Variables for Cycle 4 Non-response Adjustment

CCC8DANY	DGC8_1A	HCC8_1	PAC8_1A	SDC8_6A
CCC8_1C	DHC8_AGE	INC8DIA5	PAC8_1J	SDC8_7A
CCC8_1L	DHC8DECF	ISC8_1	PY_8DH1	SEX
CCC8_1N	DHC8_OWN	NU_8_1B	RAC8F1	SSC8DEMO
CCC8_1V	FIC8F1	PAC8DFD	RPC8_2	SSC8DSOC
DGC8F1	GE38DURB	PAC8DLEI	SDC8_4A	TWC8_5

To adjust for longitudinal members who did not respond in Cycle 4, the following adjustment is applied to the weight of respondents:

$$\frac{\text{Sum of weights for Cycles 1, 2 and 3 responding longitudinal members}}{\text{Sum of weights for Cycles 1 to 4 responding longitudinal members}}$$

This adjustment is performed within each weighting class. Records for which the panel member was deceased in Cycle 2 or 3 or institutionalized since Cycle 2 or 3 are not part of this adjustment.

Adjustment 4: Adjustment for Cycle 5 (2002/2003) Non-response

The 13,596[10] records with a full response after four cycles are taken as the starting point. Once again, records for which the panel member was deceased in Cycle 2, 3 or 4 or institutionalized since Cycle 2, 3 or 4 are treated differently from the rest. For these records, no non-response adjustment is made since their weight in Cycle 2, 3 or 4 has been already adjusted to reflect the fact that some of the Cycle 2, Cycle 3 or Cycle 4 non-respondents may have in fact been deceased or institutionalized.

Adjusting for non-response was done using the weighting class approach. Separate weighting classes were created for each design province—i.e., the 1994/1995 province of residence. When adjusting for non-response in Cycle 5, only the Cycle 4 household characteristics as well as personal characteristics of the longitudinal member were considered. As for Cycle 4, characteristics related to the sampling design of the survey or to the sampling weight were considered in

10. When the Cycle 4 data were released, there were 13,582 records in the longitudinal full subset. With the information from Cycle 8, we went back to previous cycles and could confirm, among other things, that some non-responses were in fact deaths and some deaths were a non-response. Some cases that had not been full/complete became so; others that had been full/complete were no longer so. These modifications raised the number of cases with a full/complete response after four cycles from 13, 582 to 13,596.

an effort to incorporate the sampling design of the survey into the analysis. Personal characteristics from the Health component were used, because they were available for all records that went into the Cycle 5 non-response adjustment.

The variables chosen by the CHAID algorithm to build the weighting classes to adjust for Cycle 5 non-response are in Table 3.E. Three Cycle 1 design variables were also used, one indicating the presence of household members under the age of 12, the other indicating the presence of household members under the age of 25, and the last one indicating the household type (adult or child). A Cycle 4 item non-response flag for income was also used. Please refer to the Data Dictionary for a complete description of the variables listed in Table 3.E.

Table 3.E: Variables for Cycle 5 Non-response Adjustment

ALC0_3	DHC0_OWN	IMM	MHC0_1J	SMC0_2
ALC0DTYP	DHC0DL12	INC0DIA5	MHC0DCH	ST_0DC4
ALC0DWKY	DHC0DLE5	ISC0_1	MHC0DDS	ST_0DC5
AM60_SHA	GE30DURB	LSC0_1	PAC0DFD	ST_0DC6
BPC0_10	GHC0_21	LSC0DPFT	PAC0DLEI	ST_0DC8
CCC0DANY	HCC0DHPC	MHC0_16	SDC0_4A	ST_0DR2
DGC0F1	HSC0DHSI	MHC0_1A	SDC0_6A	ST_0DW3
DHC0_AGE	HWC0DSW	MHC0_1F	SEX	ST_0DW6

To adjust for longitudinal members who did not respond in Cycle 5, the following adjustment is applied to the weight of respondents:

$$\frac{\text{Sum of weights for Cycles 1 to 4 responding longitudinal members}}{\text{Sum of weights for Cycles 1 to 5 responding longitudinal members}}$$

This adjustment is performed at the weighting class level. Records for which the panel member was deceased in Cycle 2, 3 or 4 or institutionalized since Cycle 2, 3 or 4 are not part of this adjustment.

Adjustment 5: Adjustment for Cycle 6 (2004/2005) Non-response

The 12,559[11] records with a full response after five cycles are taken as the starting point. Once again, records for which the panel member was deceased in Cycle 2, 3, 4 or 5 or institutionalized since Cycle 2, 3, 4 or 5 are treated differently from the rest. For these records, no non-response adjustment is made since their weight in Cycle 2, 3, 4 or 5 has been already adjusted to reflect that some of the Cycle 2, 3, 4 or 5 non-respondents may have in fact been deceased or institutionalized.

11. When the Cycle 5 data were released, there were 12,546 records in the longitudinal full subset. With the information from Cycle 8, we went back to previous cycles and could confirm, among other things, that some non-responses were in fact deaths and some deaths were a non-response. Some cases that had not been full/complete became so; others that had been full/complete were no longer so. These modifications raised the number of cases with a full/complete response after five cycles from 12,546 to 12,559.

Adjusting for non-response was done using the weighting class approach. Separate weighting classes were created for each design province—i.e., the 1994/1995 province of residence. When adjusting for non-response in Cycle 6, only the Cycle 5 household characteristics and the personal characteristics of the longitudinal member were considered. As for Cycle 5, characteristics related to the sampling design of the survey or to the sampling weight were considered, in an effort to incorporate the sampling design of the survey into the analysis. Personal characteristics from the Health component were used, because they were available for all records that went into the Cycle 6 non-response adjustment.

The variables chosen by the CHAID algorithm to build the weighting classes to adjust for Cycle 6 non-response are in Table 3.F. Two variables from the Cycle 1 sample design were used: a flag indicating the presence of household members less than 12 years of age, and a flag indicating the presence of household members less than 25 years of age. A Cycle 5 item non-response flag for income was also used. Please refer to the Data Dictionary for a complete description of the variables listed in Table 3.F.

Table 3.F: Variables for Cycle 6 Non-response Adjustment

ALC2_2	DHC2DL12	LSC2_1	PAC2_3E	SSC2DTNG
ALC2_3	EDC2_4	LSC2_21	PAC2_3F	ST_2DC5
ALC2DWKY	EDC2D2	LSC2DPFT	PAC2DFD	ST_2DC6
CCC2_1C	GE32DURB	MHC2_1A	PAC2DFM	ST_2DC7
CCC2_1F	HSC2DEMO	MHC2_1C	RAC2_1C	ST_2DC8
CCC2_1L	HWC2DISW	MHC2_1F	RAC2_6D	ST_2DC9
CCC2DANY	HCC2DMDC	MHC2_1G	SEX	ST_2DW1
CCC2DNUM	IMM	PAC2_1F	SMC2_2	ST_2DW2
DHC2_AGE	INC2_1A	PAC2_3A	SMC2DTYP	ST_2DW5
DHC2_OWN	INC2DHH	PAC2_3B	SMC2DYRS	ST_2DW6
DHC2DHSZ	INC2DIA5			

To adjust for longitudinal members who did not respond in Cycle 6, the following adjustment is applied to the weight of respondents:

$$\frac{\text{Sum of weights for Cycles 1 to 5 responding longitudinal members}}{\text{Sum of weights for Cycles 1 to 6 responding longitudinal members}}$$

This adjustment is performed at the weighting class level. Records for which the panel member was deceased in Cycle 2, 3, 4 or 5 or institutionalized since Cycle 2, 3, 4 or 5 are not part of this adjustment.

Adjustment 6: Adjustment for Cycle 7 (2006/2007) Non-response

The 11,619[12] records with a full response after six cycles are taken as the starting point. Once again, records for which the panel member was deceased in Cycle 2, 3, 4, 5 or 6 or institutionalized since Cycle 2, 3, 4, 5 or 6 are treated differently from the rest. For these records, no non-response adjustment is made, because their weight in Cycle 2, 3, 4, 5 or 6 has been already adjusted to reflect that some of the Cycle 2, 3, 4, 5 or 6 non-respondents may have in fact been deceased or institutionalized.

Adjusting for non-response was done using the weighting class approach. Separate weighting classes were created for each design province—i.e., the 1994/1995 province of residence. When adjusting for non-response in Cycle 7, only the Cycle 6 household characteristics and the personal characteristics of the longitudinal member were considered. As for Cycle 6, characteristics related to the sampling design of the survey or to the sampling weight were considered in an effort to incorporate the sampling design of the survey into the analysis. Personal characteristics from the Health component were used, because they were available for all records that went into the Cycle 7 non-response adjustment.

The variables chosen by the CHAID algorithm to build the weighting classes to adjust for Cycle 7 non-response are in Table 3.G. Two variables from the Cycle 1 sample design were used: a flag indicating the presence of household members less than 25 years of age, and a flag indicating the household type (adult or child). A Cycle 6 age group variable was also used. Please refer to the Data Dictionary for a complete description of the variables listed in Table 3.G.

Table 3.G: Variables for Cycle 7 Non-response Adjustment

ALCA_2	DHCA_MAR	INCADHH	LSCA_21	SSCADAFF
ALCA_3	EDCA_4	INCADIAG5	PACADFD	STCADW1
DHCADECF	GHCA_2	INCA_3B	SDCA_6A	STCADW6
DHCADLVG	IMM	LSCADPFT	SMCA_2	STCADW7
DHCA_AGE				

To adjust for longitudinal members who did not respond in Cycle 7, the following adjustment is applied to the weight of respondents:

$$\frac{\text{Sum of weights for Cycles 1 to 6 responding longitudinal members}}{\text{Sum of weights for Cycles 1 to 7 responding longitudinal members}}$$

12. When the Cycle 6 data were released, there were 11,593 records in the longitudinal full subset. With the information from Cycle 8, we went back to previous cycles and could confirm, among other things, that some non-responses were in fact deaths and some deaths were a non-response. Some cases that had not been full/complete became so; others that had been full/complete were no longer so. These modifications raised the number of cases with a full/complete response after six cycles from 11,593 to 11,619.

This adjustment is performed at the weighting class level. Records for which the panel member was deceased in Cycle 2, 3, 4, 5 or 6, or institutionalized since Cycle 2, 3, 4, 5 or 6, are not part of this adjustment.

Adjustment 7: Adjustment for Cycle 8 (2008/2009) Non-response

The 10,992[13] records with a full response after seven cycles are taken as the starting point. Again, records for which the panel member was deceased in Cycle 2, 3, 4, 5, 6 or 7, or institutionalized since Cycle 2, 3, 4, 5, 6 or 7, are treated differently from the rest. For these records, no non-response adjustment is made, because their weight in Cycle 2, 3, 4, 5, 6 or 7 has been already adjusted to reflect the fact that some of the Cycle 2, 3, 4, 5, 6 or 7 non-respondents may have in fact been deceased or institutionalized.

Adjusting for non-response was done using the weighting class approach, as with the previous cycles. Separate weighting classes were created for each design province—i.e., the 1994/1995 province of residence. When adjusting for non-response in Cycle 8, only the Cycle 7 household characteristics and the personal characteristics of the longitudinal member were considered. As for Cycle 7, characteristics related to the sampling design of the survey or to the sampling weight were considered in an effort to incorporate the sampling design of the survey into the analysis. Personal characteristics from the Health component were used, because they were available for all records that went into the Cycle 8 non-response adjustment.

The variables chosen by the CHAID algorithm to build the weighting classes to adjust for Cycle 8 non-response are in Table 3.H. One variable from the Cycle 1 sample design was used: a flag indicating the presence of household members less than 25 years of age. A Cycle 7 age group variable was also used. Please refer to the Data Dictionary for a complete description of the variables listed in Table 3.H.

Table 3.H: Variables for Cycle 8 Non-response Adjustment

ALCB_3	DHCB_MAR	INCBDHH	IMM	DHCB_OWN
DHCB_AGE	EDCBD2	INCBDPER	PACBDFD	PACB_1F
DHCBDECF	GHCB_2	GE3BDURB	SDCB_6A	SDCB_6B
MHCB_1F	CCCBDANY	LSCBDPFT	SMCB_2	HCCBDMDC
SEX				

13. When the Cycle 7 data were released, there were 10,992 records in the longitudinal full subset. With the information from Cycle 8, we went back to previous cycles and could confirm, among other things, that some non-responses were in fact deaths and some deaths were a non-response. Some cases that had not been full/complete became so; others that had been full/complete were no longer so. However, these modifications did not change the number of cases with a full/complete response after seven cycles: 10,992.

To adjust for longitudinal members who did not respond in Cycle 8, the following adjustment is applied to the weight of respondents:

$$\frac{\text{Sum of weights for Cycles 1 to 7 responding longitudinal members}}{\text{Sum of weights for Cycles 1 to 8 responding longitudinal members}}$$

This adjustment is performed at the weighting class level. Records for which the panel member was deceased in Cycle 2, 3, 4, 5, 6 or 7, or institutionalized since Cycle 2, 3, 4, 5, 6 or 7, are not part of this adjustment.

Adjustment 8: Post-stratification Adjustment

The weight of the units that are part of the full subset was post-stratified to the 1994/1995 population estimates based on 1996 Census counts by age group (0 to 11, 12 to 24, 25 to 44, 45 to 64, 65 and older) and sex within each province. This is done to ensure that the 1994/1995 population is accurately represented in any estimates produced from the longitudinal file. This adjustment is given by:

$$\frac{\text{Population estimate in a province/age/sex category}}{\text{Sum of weights of Cycles 1 to 8 responding longitudinal members in a province/age/sex category}}$$

The final longitudinal weight **WT6CLF** is calculated by taking the Cycle 1 stripped weight and multiplying that value by adjustments 1 to 8.

8.1.2.3 Longitudinal Square Share Weight (WT6CSLS)

Creation of the share square subset was started after Cycle 5. This subset includes the panel members who agreed to share the information provided from all interviews conducted as part of the NPHS with provincial ministries of health, Health Canada and the Public Health Agency of Canada. As these partners only receive the records of these sharers, a special weight, WT6CSLS, must be derived so that estimates computed from this subset correctly represent the total population.

To calculate this weight, the following two adjustments were applied to the Cycle 1 stripped weight: a non-response adjustment (do not agree to share) and a post-stratification adjustment.

Adjustment 1: Non-response Adjustment (do not agree to share)

As for the full weight computation (section 8.1.2.2), adjusting for the non-response was done using the weighting class approach. Weighting classes consist of groupings of panel members who share the same propensity to agree to share their data. The household characteristics and the personal characteristics of the longitudinal panel members from each of the eight cycles were considered to define membership in the weighting classes. The CHAID algorithm was used to determine the weighting classes: a minimum of 30 units per weighting class was used in order to produce more stable adjustments. Separate weighting classes were created for each province, i.e., the 1994/1995 province of residence.

The variables chosen by the CHAID algorithm to build weighting classes for the non-response adjustment are listed in Table 3.I. Variables were used that identified if the panel member had a status of "complete," "deceased," "institutionalized," "partial response" or "non-response" to Cycles 1 to 7 or 8. Also used was flag indicating whether the panel member was part of the full subset. Two variables from the Cycle 1 sample design were used: a flag indicating the presence of household members under 25 years of age, and a flag indicating the type of respondent (adult or child). A flag identifying all the respondents who changed provinces since Cycle 1 was also used in some of the weighting classes. Please refer to the Data Dictionary for a complete description of the variables listed in Table 3.I.

Table 3.I: Variables for the Non-response Adjustment

DHC4_MAR	DHC8_MAR	DHCB_OWN	EDCBD2	DHCCDECF
DHC2_MAR	DHCC_MAR	DHC8_OWN	EDC2D2	DHCBDECF
GE34DURB	GHCC_2	EDCCD2	EDC8D2	DHC0DECF
GHC2_2	HCCCDMDC	HCC0DMDC	INC4DHH	HCC4DMDC
INCCDPER	INC8DPER	INC0DPER	INC40HH	INCADHH
INCCAPER	LSCCDPFT	LSCC0PFT	SDC4_6A	LFC4DHMN
DHC4_AGE				

To adjust for longitudinal members who refused to share their data in Cycle 8, the following adjustment is applied to the weight of the sharers:

$$\frac{\textit{Sum of weights for all longitudinal members}}{\textit{Sum of weights for longitudinal members agreeing to share}}$$

This adjustment is performed within each weighting class.

Adjustment 2: Post-stratification Adjustment

The weight of the units that are part of the square share subset was post-stratified to the 1994/1995 population estimates based on 1996 Census counts by age group (0 to 11, 12 to 24, 25 to 44, 45 to 64, 65 and older) and sex within each province. This is done to ensure that the 1994/1995 population is accurately represented in any estimates produced from the longitudinal file. This adjustment is given by

$$\frac{\textit{Population estimate in a province/age/sex category}}{\textit{Sum of weights for longitudinal members agreeing to share in Cycle 8 in a province/age/sex category}}$$

The final weight for the square share subset, **WT6CSLS**, is calculated by taking the Cycle 1 stripped weight and multiplying this value by adjustments 1 and 2.

8.1.2.4 Longitudinal Full Share Weight (WT6CSLF)

As for the square share subset, the full share subset includes the panel members who agreed to share the information provided from all interviews conducted as part of the NPHS with provincial ministries of health, Health Canada and the Public Health Agency of Canada, but only those who had a full response in Cycles 1 to 8. As these partners only receive the records of these sharers, a special weight must be derived so that estimates computed from this subset correctly represent the total population.

A single adjustment is made to the longitudinal full weight to create the full share weight. This adjustment is given by

$$\frac{\text{Sum of weights for Cycles 1 to 8 responding longitudinal members in a province/longitudinal pattern/age-sex category}}{\text{Sum of weights for Cycles 1 to 8 responding longitudinal members who agreed to share, in a province/longitudinal pattern/age-sex category}}$$

In Cycles 3 to 8, a few of the original longitudinal response patterns were collapsed in order to produce adjustments that were more stable. The grouping was done for a few province/age–sex categories that had few observations in some of the longitudinal patterns representing the deceased or institutionalized persons. In each case, the problematic response pattern was grouped with another longitudinal pattern in the same province/age–sex category, so that the sum of the weights would still give the correct population counts. The final longitudinal share weight **WT6CSLF** is obtained by multiplying the longitudinal full weight, WT6CLF, by this adjustment. Because this adjustment is done with respect to the post-stratification classes, no additional post-stratification is necessary.

9. Data Quality

Data quality is an important aspect of any survey. Examining data quality enables verification of the reliability and accuracy of the data collected, and helps determine what steps should be taken to improve data quality in future cycles.

The survey produces estimates based on information collected from a sample of individuals. Somewhat different estimates might have been obtained if a complete census had been taken using the same questionnaire, interviewers, supervisors, processing methods, etc. as those used in the survey. The difference between the estimates obtained from the sample, and those resulting from a complete count taken under similar conditions is called the 'sampling error' of the estimate.

Errors that are unrelated to sampling can occur at almost every phase of a survey operation. Interviewers may misunderstand instructions, respondents may misunderstand questions, the answers may be incorrectly entered, or errors may be introduced in the processing and tabulation of the data. These are all examples of 'non-sampling errors'.

9.1 Sampling Errors

Given that estimates from a sample survey are unavoidably subject to sampling error, sound statistical practice calls for researchers to provide users with some indication of the magnitude of this sampling error. The basis for measuring the potential size of sampling errors is the standard deviation of the estimates derived from survey data. However, because of the large variety of estimates that can be produced from a survey, the standard deviation of an estimate is usually expressed relative to the estimate to which it pertains. This resulting measure, known as the coefficient of variation (CV) of an estimate, is obtained by dividing the standard deviation of the estimate by the estimate itself: it is expressed as a percentage of the estimate.

For example, suppose hypothetically that one estimates that 25% of Canadians aged 12 and older saw improvement in their general health between Cycle 1 and Cycle 2 of the survey, and that this estimate is found to have a standard deviation of .003. Then the CV of the estimate is calculated as

$$(.003/.25) \times 100\% = 1.20\%.$$

Statistics Canada commonly uses CV results to verify the quality of statistical estimates produced when analysing data, and strongly urges users producing estimates from National Population Health Survey (NPHS) data files to also do so. For guidelines on how to interpret CV results, see the table at the end of Section 10.4.

9.1.1 Bootstrap Method for Variance Estimation

To determine the quality of the estimate and to calculate the CV, the standard deviation must be calculated. Confidence intervals and many statistical tests also require the standard deviation of the estimate.

The NPHS uses a multi-stage survey design, which means that no simple formula can be used to calculate variance estimates. Therefore, an approximate method is needed. The bootstrap method is used because the sample design information

needs to be taken into account when calculating variance estimates. The bootstrap method does this, and with the use of the Bootvar program, remains a fairly easy method to use.

The bootstrap re-sampling method used in the NPHS involves the selection of simple random samples known as replicates, and the calculation of the variation in the estimates from replicate to replicate. In each stratum, a simple random sample of (n-1) of the n clusters is selected with replacement to form a replicate. Note that since the selection is with replacement, a cluster may be chosen more than once. In each replicate, the survey weight for each record in the (n-1) selected clusters is recalculated. These weights are then post-stratified according to demographic information in the same way as the sampling design weights in order to obtain the final bootstrap weights.

The entire process (selecting simple random samples, recalculating and post-stratifying weights for each stratum) is repeated B times, where B is large. The NPHS typically uses B=500, to produce 500 bootstrap weights. To obtain the bootstrap variance estimator, the point estimate for each of the B samples must be calculated. The standard deviation of these estimates is the bootstrap variance estimator. Statistics Canada has developed a program that can perform all of these calculations for the user: the Bootvar program. For more information on Bootstrap weights, please refer to Section 11.2.

The Bootvar program is available in both SAS and SPSS formats. It is made up of macros that compute variances for totals, ratios, differences between ratios and for linear and logistic regression.

The Bootvar program is provided with bootstrap weights and a document explaining how to modify and use the program to suit users' needs.

9.2 Non-sampling Errors

Considerable time and effort was made to reduce non-sampling errors in the NPHS. Quality assurance measures were implemented at each step of data collection and processing to monitor data quality. These measures included the use of highly skilled interviewers, extensively trained on the survey procedures, concepts and questionnaire, as well as the observation of interviewers to detect problems and give solution if needed. Testing of the CAI application and field tests were also essential procedures to ensure that data collection errors were minimized.

A major source of non-sampling errors in surveys is the effect of non-response on the survey results. The extent of non-response varies from partial non-response (failure to answer just one or some questions) to total non-response. Partial non-response to NPHS is minimal; once the questionnaire is started, it tends to be completed with very little non-response. In most cases, partial non-response to the survey occurred when the respondent refused to answer a question, could not recall the requested information, or could not provide personal or proxy information. Total non-response occurred because it was impossible to trace or reach the respondent, no member of the household was able to provide the information, or the respondent refused to participate in the survey. Total non-response was handled by adjusting the weight of persons who responded to the survey to

compensate for those who did not respond. See Section 8.1.2 for details of the weight adjustment for non-response.

This section covers different aspects of non-response. Discussed first is overall non-response, where response rates from each cycle are presented. Refusals, panel member untraced, and attrition are discussed. Finally, item non-response is briefly examined.

9.2.1 Response Rates

This chapter presents the response rates and describes how they are computed. The calculation of Cycle 1 response rates is not the same as the calculation of the response rates for the other cycles. Cycle 1 response rates are based on the 20,095 in-scope persons selected to form the panel while response rates for subsequent cycles are based on the 17,276 individuals who form the longitudinal panel. Another important difference: for the first three cycles, the selected-person response rate is calculated both for the General component (H05) and for the Health component (H06) (see Section 4.1). Since the survey became purely longitudinal in Cycle 4 and there was no longer a distinction between these two components, there is only one longitudinal panel response rate since Cycle 4.

9.2.1.1 Cycle 1 (1994/1995) Response Rates

Cycle 1 response rates are based on the 20,095 in-scope persons selected to form the panel. Consequently, persons who were part of the 3,165 out-of-scope households (status code = 017, 018, 023, 024)[14] and the 2,983 excluded households by the rejective method (see Section 5.1.2) were excluded from the panel and from the calculations of the Cycle 1 response rates.

Selected-person Response Rate for H05

$$\frac{\text{Number of selected persons responding to the H05 component}}{\text{All in-scope selected persons}}$$

The selected-person response rate for the H05 component at the Canada level for the NPHS was **86.0%**. At the provincial level, this rate varied from 80.7% in Ontario to 91.0% in Alberta.

Selected-person Response Rate for H06

$$\frac{\text{Number of selected persons responding to the H06 component}}{\text{All in-scope selected persons}}$$

The selected-person response rate for the H06 component was **83.6%** at the Canada level, and ranged from 77.8% in Ontario to 89.1% in Alberta.

14. 017 = Other ineligible dwelling (e.g., embassy).
018 = Rejected household.
023 = Under construction or demolished.
024 = Vacant dwelling.

Relevant information for the calculation of response rates is given in Table 4.A; response rates at the national and provincial level are presented in Table 4.B.

Table 4.A: Relevant Information for Calculation of Response Rates for Cycle 1

Cycle 1 (1994/1995)				
Number of in-scope selected persons (1) = (2)+(4) or (3)+(5)	**Number of respondents**		**Number of non-respondents**	
	H05 (2)	**H06 (3)**	**H05 (4)**	**H06 (5)**
20,095	17,276	16,794	2,819	3,301

Table 4.B: Cycle 1 Response Rates

Province	**Cycle 1 Response Rates (1994/1995) (%)**	
	H05 (2) / (1)	**H06 (3) / (1)**
Newfoundland and Labrador*	89.3	86.9
Prince Edward Island	87.6	84.9
Nova Scotia	85.4	82.1
New Brunswick	88.1	86.0
Quebec	87.5	85.6
Ontario	80.7	77.8
Manitoba	89.5	87.0
Saskatchewan	88.5	86.9
Alberta	91.0	89.1
British Columbia	85.2	82.8
Canada	**86.0**	**83.6**

* Labrador was not part of Newfoundland when sample was selected.

9.2.1.2 Cycle 2 (1996/1997) and Cycle 3 (1998/1999) Response Rates

All Cycle 2 and Cycle 3 response rates are based on the 17,276 individuals who form the longitudinal panel. Persons with a status “deceased” or “institutionalized’ are counted as a response for longitudinal purposes. However, persons with a status “partial response” are counted as a non-response (see Section 7.6). No panel members are classified as out of scope.

Panel response rate for H05

$$\frac{\text{Number of panel members responding to the H05 component or who have died or been institutionalized}}{\text{Number of panel members}}$$

At the Canada level, the panel response rate for the H05 component was **93.6%** in Cycle 2 and **88.9%** in Cycle 3. At the provincial level, this rate varied from 90.4% in British Columbia to 96.1% in Newfoundland and Labrador in Cycle 2 and from 84.2% in British Columbia to 92.5% in Newfoundland and Labrador in Cycle 3.

Panel response rate for H06

$$\frac{\text{Number of panel members responding to the H06 component or who have died or been institutionalized}}{\text{Number of panel members}}$$

At the Canada level, the panel response rate for the H06 component was **92.8%** in Cycle 2 and **88.3%** in Cycle 3. At the provincial level, this rate varied from 89.8% in British Columbia to 94.8% in Newfoundland and Labrador in Cycle 2 and from 83.9% in British Columbia to 92.0% in Newfoundland and Labrador in Cycle 3.

Relevant information for the calculation of response rates is given in Table 4.C; response rates at the national and provincial level are presented in Table 4.D.

Table 4.C: Relevant Information for Calculation of Response Rates for Cycles 2 and 3

Cycle	Number of Panel Members[15]					
	Deceased *(1)*	Institutionalized *(2)*	Complete		Non-respondent	
			H05 *(3)*	H06 *(4)*	H05 *(5)*	H06 *(6)*
2	289	62	15,819	15,687	1,112	1,238
3	615	114	14,647	14,532	1,912	2,015

15. With the information from Cycle 8, we went back to previous cycles and could confirm, among other things, that some non-responses were in fact deaths and some deaths were a non-response. Some cases that had not been full/complete became so; others that had been full/complete were no longer so. Numbers shown in this table show the situation after the modifications.

Table 4.D: Panel Response Rate for Cycles 2 and 3

Province	Response rate (%)			
	Cycle 2 (1996/1997)		Cycle 3 (1998/1999)	
	H05 (1) + (2) +(3) / 17 276	H06 (1) + (2) +(4) / 17 276	H05 (1) + (2) +(3) / 17 276	H06 (1) + (2) +(4) / 17 276
Newfoundland and Labrador*	96.1	94.8	92.5	92.0
Prince Edward Island	95.0	94.3	91.3	90.7
Nova Scotia	94.6	94.0	90.1	89.4
New Brunswick	94.8	94.3	89.5	89.2
Quebec	95.1	94.1	89.3	88.3
Ontario	92.1	91.4	87.6	86.9
Manitoba	95.4	94.4	91.1	90.6
Saskatchewan	94.7	94.0	91.0	91.0
Alberta	91.8	91.5	88.7	88.4
British Columbia	90.4	89.8	84.2	83.9
Canada	**93.6**	**92.8**	**88.9**	**88.3**

* Labrador was not part of Newfoundland when sample was selected.

9.2.1.3 Cycle 4 (2000/2001), Cycle 5 (2002/2003), Cycle 6 (2004/2005), Cycle 7 (2006/2007) and Cycle 8 Response Rates

As for Cycles 2 and 3, the Cycles 4, 5, 6, 7 and 8 longitudinal response rates are based on the 17,276 members of the longitudinal panel. Persons with a status “deceased” or “institutionalized’ are counted as a response for longitudinal purposes. However, those with a “partial response” status are counted as a non-response (see Section 7.6). No panel members are classified as out of scope. As of Cycle 4, the NPHS became purely longitudinal, and no longer distinguishes the H05 questionnaire from the H06 questionnaire. Only one response rate is calculated and is equivalent to the H06.

Response rate

$$\frac{\text{Number of panel members responding or who have died or been institutionalized}}{\text{Number of longitudinal panel members}}$$

At the Canada level, the panel member response rate was **84.9%** in Cycle 4, **80.8%** in Cycle 5, **77.6%** in Cycle 6 and **77.0**% in Cycle 7. In Cycle 8, the response rate was **70.7%.**

Relevant information for the calculation of response rates is given in Table 4.E; response rates at the national and provincial levels are presented in Table 4.F.

Table 4.E: Relevant Information for Calculation of Response Rates for Cycles 4 to 8

Cycle	Number of Panel Members[16]			
	Deceased (1)	Institutionalized (2)	Complete (3)	Non- respondent (4)
4	976	133	13,560	2,607
5	1,311	161	12,483	3,321
6	1,682	144	11,588	3,862
7	2,071	148	11,119	3,938
8	2,373	163	9,680	5,060

Table 4.F: Panel Response Rate for Cycles 4, 5, 6, 7 and 8

Province	Response Rate (%) = $\frac{(1) + (2) + (3)}{17,276}$				
	Cycle 4	Cycle 5	Cycle 6	Cycle 7	Cycle 8
Newfoundland and Labrador*	89.2	82.6	80.9	83,1	72,4
Prince Edward Island	88.2	84.1	80.6	82,7	74,5
Nova Scotia	87.0	81.6	81.0	82,6	75,8
New Brunswick	83.8	79.4	75.8	78,6	68,5
Quebec	85.5	79.8	75.2	77,4	73,7
Ontario	82.3	80.1	75.0	73,0	67,1
Manitoba	89.0	82.7	81.6	78,4	71,7
Saskatchewan	90.9	84.8	82.9	81,1	74,6
Alberta	83.3	80.3	79.3	77,7	70,0
British Columbia	80.8	77.9	76.5	73,7	67,5
Canada	**84.9**	**80.8**	**77.6**	**77,2**	**70,7**

* Labrador was not part of Newfoundland when the sample was selected.

16. With the information from Cycle 8, we went back to previous cycles and could confirm, among other things, that some non-responses were in fact deaths and some deaths were a non-response. Some cases that had not been full/complete became so; others that had been full/complete were no longer so. Numbers shown in this table show the situation after the modifications.

9.2.2 Refusal Rates

In a longitudinal survey, non-response to one or more cycles is quite costly. In a sense, it breaks the chronological sequence of information available on a respondent, which may make analysis of their data more complex. Worse yet, chronic non-response to the survey reduces the sample size available for analysis, diminishing the potential for statistically significant results. Consequently, many efforts have been made in the NPHS to minimize non-response.

Despite all efforts made to convert refusals (see Section 6.4), they remain the largest source of non-response for the NPHS. Although the intention is to follow all 17,276 panel members over time, not all records (panel members) are sent out for collection each cycle, such as the categorical refusals. When the panel member has been confirmed dead through a match to the mortality files or the year of death of the panel member is known, the case is considered complete for the duration of the survey, and is no longer sent out for collection.

Two different refusal rates for each cycle can be calculated—one based only on those panel members that were sent out for collection, the other based on all 17,276 panel members. Table 4.G shows both these rates for Cycles 2 to 8: both refusal rates increased with each cycle, but seem to have stabilized over time. However, efforts to convert refusals are worthwhile because some panel members that had refused for a few consecutive cycles have ended up participating in subsequent cycles of the survey.

Table 4.G: Refusal Rates by Cycle[17]

Cycle	Panel members that went out	Refusals during collection	Refusal rate based on panel members sent out	Refusals that were not sent out	Total refusals	Refusal rate based on all 17,276 panel members
	number	number	%	number	number	%
2	17,266	538	3.1	1	539	3.1
3	16,582	602	3.6	460	1,062	6.1
4	16,186	1,014	6.3	514	1,528	8.8
5	15,616	1,300	8.3	664	1,964	11.4
6	14,743	1,071	7.3	1,213	2,284	13.1
7	13,857	842	6.1	1,415	2,257	13.2
8	13,446	1,181	8.8	1,423	2,604	15.1

17. With the information from Cycle 8, we went back to previous cycles and could confirm, among other things, that some non-responses were in fact deaths and some deaths were a non-response. Numbers presented in this table show the situation after the modifications.

9.2.3 Unable-to-trace Rates

After refusal, the failure to trace a longitudinal panel member is the second greatest source of non-response for the NPHS. Despite numerous efforts from interviewers (discussed in Section 6.4), the cumulative unable-to-trace rate is rising with the passing cycles, despite many attempts that were put in place to keep this rate as low as possible. Two different unable-to-trace rates for each cycle can be calculated—one based only on those panel members that were sent out for collection, the other based on all 17,276 panel members. Table 4.H shows both these rates for Cycles 2 to 8: both unable-to-trace rates rose with each cycle.

Table 4.H: Unable-to-trace Rates by Cycle[18]

Cycle	Panel members that went out	Unable to trace during collection	Unable to trace rate based on panel members sent out	Unable to trace that were not sent out	Total unable to trace	Unable to trace rate based on all 17,276 panel members
	number	number	%	number	number	%
2	17,266	295	1.7	0	295	1.7
3	16,582	359	2.2	0	359	2.1
4	16,186	501	3.1	0	500	2.9
5	15,616	698	4.5	0	698	4.0
6	14,743	868	5.9	2	870	5.0
7	13,857	684	5.0	244	928	5.4
8	13,446	1,013	7.5	276	1,289	7.5

9.2.4 Attrition Rates

In a longitudinal survey, attrition is a loss in sample size owing to non-respondents. For the first five cycles, the attrition of the NPHS sample was defined only by whether or not a panel member was part of the full subset. Therefore, when a non-response was observed for a panel member, it was considered part of attrition. Defining attrition in this way enabled users to better understand this subset of data, but it painted a pessimistic portrait of the sample's actual state. Attrition can be defined in different ways depending on the subset of data and the method of analysis used. This section presents the attrition based on the full subset of data. Another way to describe attrition is also discussed in Section 13, where the method called 'cycle twinning' is explained. This method reduces the impact of attrition on the data. For more information on the different subsets of data, see Section 7.7.

18. With the information from Cycle 8, we went back to previous cycles and could confirm, among other things, that some non-responses were in fact deaths and some deaths were a non-response. Numbers presented in this table show the situation after the modifications.

9.2.4.1 Attrition Rates Based on the Longitudinal Full Subset

Here, attrition has been defined by whether a panel member is part of the full subset. Therefore, when a non-response is observed for a panel member, it is considered attrition. Two different attrition rates are calculated: one showing the attrition rate between two consecutive cycles, the other showing the cumulative attrition rate based on the original sample. Both of these rates are calculated using the number of individuals found in the full subset of respondents.

Relevant information for calculation of attrition rates

Number of longitudinal panel members	17,276
Number of individuals in the Cycle 2 Full subset	15,672[19]
Number of individuals in the Cycle 3 Full subset	14,630[20]
Number of individuals in the Cycle 4 Full subset	13,596[21]
Number of individuals in the Cycle 5 Full subset	12,559[22]
Number of individuals in the Cycle 6 Full subset	11,619[23]
Number of individuals in the Cycle 7 Full subset	10,992
Number of individuals in the Cycle 8 Full subset	9,982

Attrition rates between two cycles

Cycle 2 (1996/1997): $\frac{17,276 - 15,672}{17,276} = \frac{1,604}{17,276} = 9.3\%$

Cycle 3 (1998/1999): $\frac{15,672 - 14,630}{15,672} = \frac{1,042}{15,672} = 6.6\%$

Cycle 4 (2000/2001): $\frac{14,630 - 13,596}{14,630} = \frac{1,034}{14,630} = 7.1\%$

Cycle 5 (2002/2003): $\frac{13,596 - 12,559}{13,596} = \frac{1,037}{13,596} = 7.6\%$

Cycle 6 (2004/2005): $\frac{12,559 - 11,619}{12,559} = \frac{940}{12,559} = 7.5\%$

Cycle 7 (2006/2005): $\frac{11,619 - 10,992}{11,619} = \frac{627}{11,619} = 5.4\%$

Cycle 8 (2008/2009): $\frac{10,992 – 9,982}{} = \frac{1,010}{} = 9.2\%$

19. When the Cycle 2 data were released, there were 15,670 records in the longitudinal full subset. With the information from Cycle 8, we went back to previous cycles and could confirm, among other things, that some non-responses were in fact deaths and some deaths were a non-response. Some cases that had not been full/complete became so; others that had been full/complete were no longer so. These modifications raised the number of cases with a full/complete response after two cycles from 15,670 to 15,672.
20. For the same reasons discussed in the previous note, after the modifications there were 14,631 records in the Cycle 3 longitudinal full subset, up from 14,619.
21. For the same reasons discussed in note 19, after the modifications there were 13,596 records in the Cycle 4 longitudinal full subset, up from 13,582.
22. For the same reasons discussed in note 19, after the modifications there were 12,559 records in the Cycle 5 longitudinal full subset, up from 12,546.
23. For the same reasons discussed in note 19, after the modifications there were 11,619 records in the Cycle 6 longitudinal full subset, up from 11,593.

10,992 10,992

Cumulative attrition rates for the longitudinal full subset

Cycle 2 (1996/1997):	17,276-15,672 / 17,276	=	1,604 / 17,276	=	9.3%
Cycle 3 (1998/1999):	17,276-14,630 / 17,276	=	2,646 / 17,276	=	15.3%
Cycle 4 (2000/2001):	17,276-13,596 / 17,276	=	3,680 / 17,276	=	21.3%
Cycle 5 (2002/2003):	17,276-12,559 / 17,276	=	4,717 / 17,276	=	27.3%
Cycle 6 (2004/2005):	17,276-11,619 / 17,276	=	5,657 / 17,276	=	32.7%
Cycle 7 (2006/2007):	17,276-10,992 / 17,276	=	6,284 / 17,276	=	36.4%
Cycle 8 (2008/2009):	17,276-9,982 / 17,276	=	7,294 / 17,276	=	42.2%

As is typically the case in longitudinal surveys, the attrition rate between Cycles 1 and 2 is considerably higher (9.3%) than those subsequently observed. The subsequent attrition rates are more constant between cycles. Cumulatively, after eight cycles, about 42% of the panel has eroded based on the full subset. Table 4.I presents the important attrition types by cycle. As noted, refusal and unable to trace are still the most substantial sources of attrition.

Table 4.I: Attrition Type by Cycle—Longitudinal Full Subset of Respondents

Attrition type	Attrition cycle by cycle (%)							Cumulative attrition (%)
	Cycle 2	Cycle 3	Cycle 4	Cycle 5	Cycle 6	Cycle 7	Cycle 8	Cycle 8
Refusal	30.6	45.5	52.2	51.6	47.7	41.0	36.3	42.7
Unable to trace	17.1	12.5	16.8	15.6	20.2	21.5	20.5	17.4
Partial response in Cycle 1	30.0	N/A	N/A	N/A	N/A	N/A	N/A	6.6
Partial response	6.9	8.9	5.9	8.3	6.9	6.5	5.7	7.0
No one home or no answers	1.9	4.7	8.2	9.4	14.0	22.4	29.7	11.4
Moved outside Canada	4.7	7.5	5.1	1.9	2.4	2.7	1.0	3.8
No interview—mental/ physical health problem	N/A	N/A	2.4	4.9	2.2	1.6	1.3	1.6
Other non-responses	8.9	20.9	9.3	8.3	6.6	4.3	5.4	9.4

9.2.5 Item Refusal and "Don't Know" Rates

9.2.5.1 Refusal and "Don't Know" Rates by Item

Item rates have been calculated from the number of refusal, "don't know" and valid values for each variable, sub-module and module in the questionnaire. Derived variables and variables with few valid responses are not presented. Valid values exclude responses coded "not applicable" or "not stated." Table 4.J shows refusal and "don't know" rates for Cycle 8 modules.

Table 4.J: Refusal and "Don't Know" Rates by Module

Module	Refusal rate (%)	"Don't Know" rate (%)
Overall	**0.23**	**0.44**
Administration	0.01	0.02
Household Record Variables (DHC)	0.03	0.05
General Health (GHC)	0.03	0.11
Sleep (SLC)	0.03	0.34
Height/Weight (HWC)	0.03	0.03
Nutrition (NU_, FV_, SK_, MK_)	0.18	0.40
Preventive Health (PHC)	0.23	0.82
Health Care Utilization (HCC)	0.03	0.18
Restriction of Activities (RAC)	0.13	0.22
Chronic Conditions (CCC)	0.01	0.77
Health Status (HSC)	0.05	0.25
Physical Activities (PAC)	0.51	0.22
UV Exposure (TUC)	0.51	0.09
Repetitive Strain (RPC)	0.01	1.41
Injuries (IJC)	0.05	0.35
Stress (STC, ST_)	0.13	0.37
Medication Use (DGC)	0.04	0.34
Smoking (SMC)	0.18	0.65
Alcohol (ALC)	0.10	0.59
Mental Health (MHC)	0.25	0.22
Social Support (SSC)	0.15	0.41
Socio-Demographic (SDC)—Language	0.46	0.13
Education (EDC)	0.06	0.03
Labour Force (LSC)	0.21	0.26
Income (INC)	1.27	2.48
Food insecurity (FI_)	0.18	0.40

This table shows that refusal rates by module are very low, between 0.01% and 1.27%. The overall refusal rate is 0.23%. The same variables or modules tended to have relatively high refusal rates in each cycle. As in the previous cycle, the income module had the highest rate (1.27%). The physical activity module had the

second highest refusal rate at 0.51%, even with UV Exposure. Deeper analysis, not shown in the table, reveals that the rates of some labour force sub-modules were among the highest refusal rates in Cycles 2 to 8, reaching 0.62% in Cycle 8. Refusal rates by variable generally fluctuated between 0.00% and 2.56%, except some variables related to income and to chronic condition (stroke) that reached refusal rates of 4.26% and 6.67%, respectively.

Rates for the "don't know" module rates were low, between 0.02% and 2.48%. The overall "don't know" rate is 0.44%. The same variables and modules tended to have relatively high "don't know" rates in each cycle. For example, the income module has a "don't know" rate of 2.48%; some sub-modules related to labour force and chronic condition have "don't know" rates of 1.90% and 4.51%, which are among the highest rates in each cycle. The repetitive strain module had the second highest rate, 1.41%. In particular, some of the sub-modules of the chronic condition had a 4.52% "don't know" rate for Cycle 8. The "don't know" rates by variable generally fluctuated between 0.00% and 7.33%: chronic condition variables had the highest rates.

9.2.5.2 Refusal and "Don't Know" Rates by Respondent

Refusal and "Don't Know" rates were also calculated to determine the percentage of questions an individual refuses to answer or answers "don't know." All of these rates are for respondents who have a "completed" status for the Cycle 8 interview. Table 4.K shows the proportion of respondents who did not refuse to answer any questions, who refused less than 1% and less than 3% of the questions asked. Also in this table are the rates for "don't know" answers to none of the questions, less than 0.5%, less than 1% and less than 5% of the questions asked.

Table 4.K: Refusal and "Don't Know" Rates by Respondent

	Refusals to 0% of questions	Refusals to less than 1% of questions	Refusals to less than 3% of questions	"Don't know" to 0% of questions	"Don't know" to less than 0.5% of questions	"Don't know" to less than 1% of questions	"Don't know" to less than 5% of questions
	%	%	%	%	%	%	%
Overall	**90.6**	**97.5**	**98.4**	**48.4**	**79.1**	**90.5**	**99.0**
Males	91.2	97.8	98.3	52.2	81.1	90.7	99.1
Females	90.1	97.3	98.4	45.3	77.5	90.4	98.9
Under 12	100.0	0.0	0.0	100.0	0.0	0.0	0.0
12 to 24	96.0	99.2	99.5	39.7	77.5	89.1	99.5
25 to 44	94.3	98.1	98.6	62.8	89.3	95.8	99.6
45 to 64	90.4	97.5	98.3	52.2	83.3	93.8	99.4
65 and older	83.3	95.9	97.5	30.0	60.7	79.6	97.4
Proxy	90.5	96.8	97.1	28.4	41.1	65.0	94.2
Non-proxy	90.6	97.5	98.4	49.3	80.7	91.6	99.2

"Don't Know" rates have more variation than the refusal rates when they are separated by sex, age group and interview type. For Cycle 8, 90.6% answered all of the questions and a little over 98% had refusal rates for less than 3% of the questions. As for "Don't Know" rates, there are 48.4% of the respondents that answered all of the questions—in other words, almost half the respondents never answered "don't know." Around 80% of the respondents answered "don't know" to less than 0.5% of the questions. For 1% and 5% of the questions asked, the "don't know' rates are 90% and 99%, respectively. Almost everyone who refuses or responds "don't know" does so for only a few questions.

10. Guidelines for Tabulation, Analysis and Release

This section outlines the guidelines that users should follow to tabulate, analyse, release or otherwise publish any data derived from the National Population Health Survey (NPHS) data. With the aid of these guidelines, users should be able to produce figures that are in close agreement with those produced by Statistics Canada. At the same time, users will be able to develop currently unpublished figures in a manner consistent with these guidelines.

10.1 Rounding Guidelines

In order that dissemination of estimates derived from NPHS data corresponds to estimates produced by Statistics Canada, users should follow these guidelines regarding the rounding of such estimates. Unrounded estimates imply greater precision than actually exists.

a) Estimates in the main body of a statistical table should be rounded to the nearest hundred units using the normal rounding technique. In normal rounding, if the first or only digit to be dropped is 0 to 4, the last digit to be retained is not changed. If the first or only digit to be dropped is 5 to 9, the last digit to be retained is raised by one. For example, in normal rounding to the nearest 100, if the last two digits are between 00 and 49, they are changed to 00 and the preceding digit (the hundreds digit) is left unchanged. If the last digits are between 50 and 99, they are changed to 00 and the preceding digit is incremented by 1.

b) Marginal subtotals and totals in statistical tables should be derived from their corresponding unrounded components, and then rounded themselves to the nearest 100 units using normal rounding.

c) Averages, proportions, rates and percentages should be computed from unrounded components (i.e., numerators and/or denominators), then rounded to one decimal using normal rounding. In normal rounding to a single digit, if the first or only digit to be dropped is 0 to 4, the last digit to be retained is not changed. If the first or only digit to be dropped is 5 to 9, the last digit to be retained is increased by 1.

d) Sums and differences of aggregates (or ratios) should be derived from their corresponding unrounded components and themselves rounded to the nearest 100 units (or the nearest one decimal) using normal rounding.

e) When technical or other limitations necessitate using a rounding technique other than normal rounding, and as a result estimates are published that differ from corresponding estimates published by Statistics Canada, users should note the reason for such differences in the publication or release document(s).

10.2 Sample Weighting Guidelines for Tabulation

The sample design used for the NPHS was not self-weighting—the sampling weights are not identical for all individuals in the sample. When producing simple estimates, including statistical tables, users must apply the proper sampling weight. If proper weights are not used, the estimates derived from the various respondent subsets cannot be considered

representative of the 1994/1995 target population, and will not correspond to estimates produced by Statistics Canada.

Also, some software packages might not allow the generation of estimates that exactly match those available from Statistics Canada because of their treatment of the weight variable.

10.2.1 Definitions of Types of Estimates: Categorical versus Quantitative

Before discussing how the NPHS data can be tabulated and analysed, it is useful to describe the two main types of point estimates of population characteristics that can be computed.

Categorical Estimates

Categorical estimates are estimates of the number or percentage of the surveyed population possessing certain characteristics or falling into some defined category. The number of individuals who quit smoking between cycles is an example of such an estimate. An estimate of the number of persons possessing a certain characteristic may also be referred to as an estimate of an aggregate.

Example of a Categorical Question

*At the present, do/does … smoke cigarettes daily, occasionally or not at all? (**SMCC_2**)*

__ Daily
__ Occasionally
__ Not at all

Quantitative Estimates

Quantitative estimates are estimates of totals or of means, medians and other measures of central tendency of quantities, based on some or all of the members of the surveyed population.

An example of a quantitative estimate is the average increase in the number of cigarettes smoked per day by daily smokers who had an increase in consumption between two cycles.

Example of Quantitative Question

*How many cigarettes do/does you/he/she smoke each day now? (**SMCC_4**)*

|_|_| Number of cigarettes

10.2.2 Tabulation of Categorical Estimates

Estimates of the number of people with a certain characteristic can be obtained by summing the final weights of all records possessing the characteristic(s) of interest. Proportions and ratios of the form $\hat{X} / \hat{Y}$ are obtained by

a) summing the final weights of records having the characteristic of interest for the numerator ($\hat{X}$),
b) summing the final weights of records having the characteristic of interest for the denominator ($\hat{Y}$), then
c) dividing the numerator estimate by the denominator estimate.

10.2.3 Tabulation of Quantitative Estimates

Estimates of sums or averages for quantitative variables can be obtained using the following three steps (only step *a)* is necessary to obtain the estimate of a sum):

a) multiplying the value of the variable of interest by the final weight and summing this quantity over all records of interest to obtain the numerator ($\hat{X}$),
b) summing the final weights of records having the characteristic of interest for the denominator ($\hat{Y}$), then
c) dividing the numerator estimate by the denominator estimate.

For example, to obtain the estimate of the average number of cigarettes smoked each day by individuals who smoke daily, first compute the numerator ($\hat{X}$) by summing the product of the value of variable **SMCC_4** and the final weight over those records with a value of "daily" to the variable **SMCC_2**. The denominator ($\hat{Y}$) is obtained by summing the final weight of those records with a value of "daily" to the variable **SMCC_2**. Divide ($\hat{X}$) by ($\hat{Y}$) to obtain the average number of cigarettes smoked each day by daily smokers.

10.3 Guidelines for Statistical Analysis

The NPHS is based on a complex sampling design, with stratification and multiple stages of selection, and unequal probabilities of selection of respondents. Using data from such complex surveys presents problems to analysts because the survey design and the selection probabilities affect the estimation and variance calculation procedures that should be used.

While many analysis procedures found in statistical packages allow weights to be used, the meaning or definition of the weight in these procedures differs from what is appropriate in a sample survey. As a result, in many cases the estimates produced by the packages are correct, but the variances that are calculated are almost meaningless.

For many analysis techniques (for example linear regression, logistic regression, or analysis of variance), a method exists that can make the application of standard packages more useful. If the weights on the records are rescaled so that the average weight is 1, then the results produced by the standard packages will be more reasonable. They still will not take into account the stratification and clustering of the sample's design, but they will take into account the unequal probabilities of selection. The rescaling can be accomplished by using in the analysis a weight equal to the final weight divided by the average of the final weights for the sampled units (people) contributing to the estimate in question.

Coefficient-of-variation (CV)tables were produced in the past for the cross-sectional data. CV tables were not created for the longitudinal files, as a very large number of possible variable combinations for analysis exist. To correctly estimate the variance, the NPHS recommends the use of the bootstrap method. With the bootstrap method, the complexity of the weighting and the survey design are incorporated into the calculation of the variance. A SAS bootstrap variance program, with accompanying documentation and examples of how to use it, has been created to facilitate calculation of the variance using the bootstrap method. The program also calculates the accompanying CV. A similar version of the program is available in SPSS. Users must learn how to use the program, as it will generate exact estimates of individual variances to assess the quality of tabulated estimates. This program is highly preferable to the scaled weights approach. Some statistical packages such as STATA have the ability to read in the stratum and cluster information to use in variance estimation, which improves the quality of the estimate but does not take into account the different adjustments applied to the weights.

10.4 Release Guidelines

Before releasing or publishing any total or proportion estimates from the master files, users must first determine the number of sampled respondents having the characteristic of interest (for example, the number of respondents who smoke when interested in the proportion of smokers for a given population). If this number is less than 10, the weighted estimate should not be released regardless of the value of the CV for this estimate. This is because the possibility of obtaining an artificially low variance is greater with a sample size less than 10. For weighted estimates based on sample sizes of 10 or more, users should determine the CV of the estimate, and follow the guidelines described in Table 5.

Table 5: Sampling Variability Guideline

Type of estimate	Coefficient of variation (in percent)	Guidelines
Acceptable	0.0 to 16.5	Estimates can be considered for general unrestricted release. Requires no special notation.
Marginal	16.6 to 33.3	Estimates can be considered for general unrestricted release but should be accompanied by a warning cautioning subsequent users of the high sampling variability associated with the estimates. Such estimates should be identified by the letter E (or a similar symbol).
Unacceptable	Greater than 33.3	Statistics Canada does not recommend releasing estimates of unacceptable quality. However, if the user chooses to do so, estimates should be flagged with the letter F (or in a similar fashion) and the following warning should accompany the estimates: "The user is advised that (specify the data) do not meet Statistics Canada's quality standards for this statistical program. Conclusions based on these data will be unreliable and most likely invalid. These data and any consequent findings should not be published. If the user chooses to publish these data or findings, then this disclaimer must be published with the data."

11. Using the Longitudinal Master File

11.1 Use of Longitudinal Weights

The Cycle 8 master file contains all 17,276 panel members (including occasional or permanent non-respondents). Four subsets of respondents have been created (Section 7.7), which correspond to a distinct set of weights (Section 8.1). Flags were created to identify records that are part of a particular subset (Table 3.A). Records that are not part of a particular subset have a flag equal to 0 and the weight variable set to blank for that particular subset. **To create the subset of interest, select those records that have the appropriate flag variable equal to 1.**

Weight WT64LS is called the 'square weight', and applies to the 17,276 members that make up the original longitudinal sample. All non-response should be taken into account for any calculation.

Weight WT6CLF is called the 'longitudinal full' weight, and applies to the 9,982 records that are included in the 'full' subset of respondents.

Weight WT6CSLS is called the 'longitudinal square share' weight, and applies to the 15,805 records that are included in the 'square share' subset of respondents.

Weight WT6CSLF is called 'longitudinal full share' weight, and applies to the 9,627 respondents that are included in the 'full share' subset of respondents.

11.2 Ensuring the Reliability of Estimates Using Bootstrap Weights

Bootstrap weights are necessary for variance estimation. Information on the bootstrap method for variance estimation can be found in Section 9.1.1. Each subset of respondents has a set of bootstrap weights associated with it. Four different sets of bootstrap weights were created for the Cycle 8 data: the square, the square share, the full and the full share. Table 6.A presents the subset of respondents with their corresponding bootstrap file name.

Table 6.A: Subsets of Respondents and Corresponding Bootstrap Weights Files

Subset of respondents	Name of the Bootstrap Weights file	Number of Respondents
Longitudinal square	B5long	17,276
Longitudinal full	B5lngf	9,982
Longitudinal square share	B5long	15,805
Longitudinal full share	B5lngf	9,627

Due to the complex sample design, users should use the supplied Bootvar program for variance calculation. The standard variance output from other statistical packages such as SAS and SPSS may grossly underestimate the variance of an estimate for this survey. **It is the responsibility of the user to ensure the quality and reliability of the estimates**

they are producing by following the guidelines in Chapter 10 and correctly calculating the variance for all estimates. Failure to do so could lead to misinterpretation of results and jeopardize the quality of the research work.

Some statistical software is capable of including the stratum and cluster information as input when performing analytical processing, which does provide a variance estimate much closer to the true variance estimate. However, these packages fail to account for the various weighting adjustments, which in some cases can affect considerably the variance estimates.

11.3 Variable-naming Convention

The National Population Health Survey (NPHS) adopted a variable-naming convention that enables users to easily use and refer to similar data from different collection periods and across survey components of the NPHS program. The following requirements were mandatory: restrict variable names to a maximum of eight characters for ease of use by analytical software products; identify the survey occasion (1994/1995, 1996/1997, 1998/1999, 2000/2001, 2002/2003, 2004/2005, 2006/2007, and 2008/2009) in the name; and allow conceptually identical variables to be easily identifiable over survey occasions. For example, conceptually identical data on smoking were collected in 1994/1995, 1996/1997, 1998/1999, 2000/2001, 2002/2003, 2004/2005, 2006/2007, and 2008/2009, and the variable names should only differ in the position that identifies the particular survey occasion in which they were collected. This convention is followed throughout the longitudinal survey, and is adopted by all NPHS surveys: the household component, the health institutions component, and previously the North component and supplements.

11.3.1 Variable Name Component Structure

Each of the eight characters in a variable name contains information about the type of data contained in the variable.

Positions 1-2:	Variable name/Questionnaire section name
Position 3:	Survey type
Position 4:	Cycle
Position 5:	Variable type
Positions 6-8:	Variable number/name from questionnaire
For example:	the variables DHC4_AGE, DHC6_AGE, DHC8_AGE, DHC0_AGE, DHC2_AGE, DHCA_AGE, DHCB_AGE and DHCC_AGE:
DH:	in the Demographic and Household content section of the questionnaire;
C:	questions which are Core content on the household survey;
4/6/8/0/2/A/B/C:	Cycle 1 (1994/1995) variable / Cycle 2 (1996/1997) variable / Cycle 3 (1998/1999) variable / Cycle 4 (2000/2001) variable / Cycle 5 (2002/2003) variable / Cycle 6 (2004/2005) variable / Cycle 7 (2006/2007) variable / Cycle 8 (2008/2009);
_:	Cycle specific focus content;

AGE: the variable name.

11.3.2 Positions 1-2: Variable Name / Questionnaire Section Name

AD	Alcohol dependence
AL	Alcohol consumption
AM	Administration (of the survey)
AP	Attitude toward parents
BF	Breast-feeding
BP	Blood pressure
CC	Chronic conditions
CO	Coping–stress—Alberta buy-in, Cycles 1 and 2
CE	Contact/exit
DG	Drug/medication use
DH	Demographics and household variables
DV	Dental visits
ED	Education
ES	Emergency services
EX	Eye examination
FH	Personal and family medical history
FI	Food insecurity—HRDC buy-in, Cycle 3
FS	Flu shots
FV	Fruit and vegetable consumption
GE	Geographic identifiers (methodology)
GH	General health
HC	Health care utilisation
HI	Health information
HH	Household
HN	Health number
HS	Health status
HV	HIV
HW	Height, weight and body image
IJ	Injuries
IN	Income
IS	Insurance
LF/LS	Labour force
MH	Mental health
MK	Milk consumption
NU	Nutrition
PA	Physical activities
PC	Physical check-up
PH	Preventive health
PR	Province
PY	Psychological resources (self-esteem, mastery, sense of coherence)
RA	Restriction of activity
RH/MB	Residential history
RP	Repetitive strain
RS	Road safety
RT	Rationality—Alberta buy-in, Cycle 1
SC	Self-care

SD Socio-demographic characteristics

Positions 1-2: Variable Name / Questionnaire Section Name (continued)

SH Sexual health
SK Soft drink consumption
SL Sleep
SM Smoking
SP Sample identifiers (methodology)
SS Social support
ST Stress
SV Health care utilisation
TA Tobacco alternatives (Health promotion 1998)
TU Tanning–UV exposure
TW Two-week disability
VS Violence and personal safety
WF Subset flags
WT Sample weights (methodology)

A few important identifying variables do not follow the naming convention, e.g., REALUKEY, PERSONID, CYCLE, SUBCYCLE, DESIGPRV, STRATUM, and REPLICAT.

Other variables are considered 'constant'. Table 6.B presents the variables that appear only once of the data file. The name of these variables does not follow the naming convention.

Table 6.B: Constant Longitudinal Variables

Longitudinal Variable Name	Concept
AOI	Age at time of immigration
COB	Country of birth
COBC	Code for country of birth
COBGC	Code for country of birth (7 groups)—grouped
COD10	Cause of death code (ICD-10)
DESIGPRV	Province of residence in 1994
DOB	Date of birth
DOD	Day of death
HWB	Birth weight
HWBG1	Birth weight—grouped
IMM	Immigration status
MOB	Month of birth
MOD	Month of death
REPLICATE	Replicate
SEX	Sex
STRATUM	Stratum
YOB	Year of birth
YOD	Year of death

Longitudinal Variable Name	Concept
YOI	Year of immigration to Canada

11.3.3 Position 3: Survey Type

A	Asthma supplement
B	Province-specific buy-in content—children's questions
C	Core questions repeated in each cycle
F	Food insecurity supplement
I	Institutions
K	Longitudinal children's questions
N	North (Yukon/Northwest Territories)
P	Province-specific buy-in content—adult questions
S	National supplement (Health Promotion Survey)
_	Cycle specific focus questions, not repeated in every cycle (e.g., food choice in Cycles 3, 5 and 7)
3	Survey administration variables for household and demographic component (H03)
5	Survey administration variables for the General component (H05)
6	Survey administration variables for the Health component (H06) (e.g., weights, agreement to share, date of interview variables)

11.3.4 Position 4: Cycle (years)

4	Cycle 1 (1994/1995)
6	Cycle 2 (1996/1997)
8	Cycle 3 (1998/1999)
0	Cycle 4 (2000/2001)
2	Cycle 5 (2002/2003)
A	Cycle 6 (2004/2005)
B	Cycle 7 (2006/2007)
C	Cycle 8 (2008/2009)
D	Cycle 9 (2010/2011)
E	Cycle 10 (2012/2013)

11.3.5 Position 5: Variable Type

Position 5	Variable type	Description
_	Collected variable	A variable that appeared directly on the questionnaire
C	Coded variable	A variable coded from one or more collected variables (e.g., North American Industry Classification System (NAICS))
D	Derived variable	A variable calculated from one or more collected or coded variables, usually calculated during head office processing (e.g., Comprehensive Health Status Measurement System [CHSMS-HUI3])
F	Flag variable	A variable calculated from one or more collected variables (like a derived variable), but usually calculated by the computer application for later use during the interview (e.g., work flag). It can also denote that a long answer was collected (e.g., restriction of activity flag)
G	Grouped variable	Collected, coded, suppressed or derived variables collapsed into groups (e.g., age groups)

L	Longitudinal derived variable	A variable calculated using variables from two or more survey cycles

11.3.6 Positions 6-8: Variable Name

In general, the last three positions follow the naming on the questionnaire. Numbers are used where possible: Q1 becomes 1. 'Mark all' questions use letters for each possible answer category: Q1 (mark all that apply) becomes 1A, 1B, 1C, etc. Demographic variables, which are used frequently by analysts, are identified by a three letter identifier, rather than by a question number. For example "Age" is DHC4_AGE in Cycle 1 (1994/1995), DHC6_AGE in Cycle 2 (1996/1997), etc. Where groups of questions with the same topic were collected in sections that had different section names on the questionnaire, position 6 is used to identify the subsection. For example, the first question on chronic stress was named STC2_C1; the first question on work stress was named STC2_W1. Another example of this occurs in the general health questions for the Health Promotion Survey. These questions were separated into three sections for inclusion in the questionnaire and the corresponding variable names reflect this, with position 6 indicating the section in which it appears.

12. Access to NPHS Data

12.1 Research Data Centres

Confidentiality concerns preclude general dissemination of longitudinal National Population Health Survey (NPHS) data in public use microdata file (PUMF) format. However, access to all the longitudinal master microdata files, including the data for Cycles 1 to 8, is available through Statistics Canada's Research Data Centres (RDC) program. The RDC program is part of an initiative by Statistics Canada, the Social Sciences and Humanities Research Council, and university consortia to help strengthen Canada's social research capacity and to support the policy research community.

RDCs provide researchers with access, in a secure university setting, to microdata from population and household surveys. The centres are staffed by Statistics Canada employees. They are operated under the provisions of the *Statistics Act* in accordance with all the confidentiality rules and are accessible only to researchers with approved projects who have been sworn in as 'deemed employees'. RDCs are located throughout the country, so researchers need not travel to Ottawa to access Statistics Canada microdata. More information is available at the Research Data Centre Program website: http://www.statcan.gc.ca/rdc-cdr/network-reseau-eng.htm

12.2 Remote Access

A second option, if researchers cannot travel to an RDC, is Health Statistics Division's Remote Access service. It offers researchers a means to develop and test their own computer programs using synthetic files that mimic the actual master files. Researchers then submit their programs to a dedicated e-mail address. The programs are run against the master microdata files on an internal secure server, outputs are vetted for confidentiality, and emailed back to the researcher. For more information on this service, please contact the Data Access and Information Services team at nphs-ensp@statcan.gc.ca.

12.3 Data Liberation Initiative

PUMFs are available for each of the first three cycles of the NPHS, providing widespread access to the cross-sectional components of the survey. The NPHS PUMFs can be accessed through the Data Liberation Initiative at participating Canadian universities and colleges. For more information, please consult the following Statistics Canada's link: http://www.statcan.gc.ca/english/Dli/dli.htm. Cycles 1, 2 and 3 NPHS PUMFs can also be purchased. To this end, please contact the Data Access and Information Services team at hd-ds@statcan.gc.ca or one of Statistics Canada's regional offices.

12.4 Analytical Reports and Tabulations

With the release of NPHS Cycle 5 data, results from the survey were presented in a free Internet Publication entitled *Healthy Today, Healthy Tomorrow? Findings from the National Population Health Survey.* The publication (Catalogue no. 82-618M) features a series of articles addressing important health issues using NPHS longitudinal data. Consult this publication at http://www.statcan.gc.ca/bsolc/english/bsolc?catno=82-618-M&CHROPG=1.

Longitudinal CANSIM tables are also available free of charge on the Statistics Canada website. They present changes, from one NPHS cycle to another one, in smoking, self-rated health, body mass index, and physical activity. To access CANSIM tables, click the link below, and then click the Summary tables link on the left side of the page.

http://www.statcan.gc.ca/cgi-bin/imdb/p2SV.pl?Function=getSurvey&SurvId=3225&SurvVer=0&InstaId=15280&InstaVer=8&SDDS=3225&lang=en&db=IMDB&adm=8&dis=2

Research articles based on the NPHS often appear in *Health Reports*, a quarterly journal produced by Health Statistics Division. This product is available as a standard printed publication (Catalogue no. 82-003-XPE) or in electronic format (Catalogue no. 82-003-XIE) on the Statistics Canada website. For more information, consult these links:
http://www.statcan.gc.ca/bsolc/english/bsolc?catno=82-003-X&CHROPG=1
http://www.statcan.gc.ca/bsolc/english/bsolc?catno=82-003-S&CHROPG=1

Custom tabulations based on the NPHS cross-sectional data are also available on a cost recovery basis. No custom tabulation is produced with the NPHS longitudinal data. For more information or estimates on costs and feasibility, contact the Health Statistics Division's Data Access and Information Services at hd-ds@statcan.gc.ca.

13. An Analytical Technique for Longitudinal Survey Data

Longitudinal surveys like the National Population Health Survey (NPHS) have the advantage of seeing their analytical potential grow over time. These longitudinal surveys enable health analysts to study events that affect an individual's health, study the effects and causes of these events, and produce incidence rates. Analysing such data highlights equally the complexity of the network of relations that exist between the health of an individual and the intensity of exposures to different risk factors. This advantage permits research of more complex research questions about health. However, certain issues reduce this analytical potential and may even pose a long term risk to longitudinal surveys.

The analysis of data from longitudinal surveys faces different issues than those from cross-sectional surveys. On the one hand, studying more complex research questions leads data users to try more complex analytical techniques, often those that are less well-known and less well-documented. On the other hand, attrition of data over time (see Section 9.2.4) is one of the key challenges in the analysis of longitudinal data. In addition to the risk of introducing a bias in the estimates, decreasing sample size due to attrition can also be problematic for both the variety of statistical analyses and their quality.

In addition, the commonest and simplest analytical approach is to use the longitudinal full subset (see Sections 7.6 and 7.7). This subset, however, is the most sensitive to attrition, which reduces the number of individuals in every cycle for this subset. After eight cycles, the cumulative attrition rate for the NPHS full subset has reached 42.2%. In other words, 57.8% of the original panel of respondents have complete responses (see Section 9.2.4) to the eight cycles of the NPHS.

The purpose of this chapter is to give users a practical analytical tool that will enable them to reduce the effects of attrition on their analyses, and thereby extend the analytical potential of the data. Consequently, mitigating the effects of attrition will increase the accuracy of the estimates and the relevance of the survey.

13.1 Cycle Twinning Approach

The cycle twinning approach, also called 'pooling of repeated observations', is an analytical approach that enables reduction of the effects of erosion. It consists of using a subset of cycles from a respondent as the unit of analysis, rather than considering all cycles from a respondent as the unit of analysis. This approach is a particular case in a type of statistical models called 'marginal models'.[24]

Although the fictitious example and the application described in this section are limited to two consecutive cycles completed by the same respondent as the unit of analysis, this approach can be adapted to many scenarios. For example, the length of exposure to different risk factors could be shorter or longer, the number of cycles (or repeated measures) could be 1 or more than 2, the dependent variables could be continuous or discrete with temporary or permanent conditions, such as death or certain chronic conditions.

24. Fitzmaurice, G.M., Laird, N.M., and Ware, J.H. (2004). *Applied Longitudinal Analysis*. New York: Wiley.

Fictitious example

Thus after eight cycles of collection, when the cycle twinning approach is used, a respondent can supply up to seven units of analysis if he or she belongs to the full subset: Cycles 1 and 2, Cycles 2 and 3, Cycles 3 and 4, Cycles 4 and 5, Cycles 5 and 6, Cycles 6 and 7 and Cycles 7 and 8 (see respondent A in Table 7.A). If a respondent did not respond to cycle 3, 4 and 7 for example, but responded to all other cycles, he or she can produce two units of analysis: Cycles 1-2 and Cycles 5-6 (see respondent B in Table 7.A). The latter respondent is not part of the Cycle 8 full subset given his non-response to Cycles 3, 4 and 7, and he or she would be automatically excluded from any analyses using this subset. However, by using the cycle twinning approach, two analytical units can come from this respondent. In order to supply at least one analytical unit, all that is required is that a respondent answers two consecutive cycles. According to Table 7.A, respondents A, B, C and D will provide 7, 2, 4 and 1 analytical units, respectively.

Table 7.A: Example of Profiles of Response from Fictitious NPHS Respondents

Respondent	Cycles with a complete response							
	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5	Cycle 6	Cycle 7	Cycle 8
A	X	X	X	X	X	X	X	X
B	X	X			X	X		X
C	X	X	X	X	X		X	
D	X		X	X			X	

According to this fictitious table, the Cycle 8 full subset would contain only one of the four respondents since the analytical unit using the full subset is respondents who answer all cycles. Thus, for this example, the attrition rate after eight cycles would be 75%.

When the cycle twinning approach is used, the analytical unit becomes a record that contains two consecutive cycles from one respondent. Thus, according to the example in Table 7.A, the potential number of analytical units is 28 (seven possible combinations of two consecutive cycles multiplied by four respondents). As calculated above, the total number of analytical units is 14 (7+2+4+1), or 50% of the potential maximum number of analytical units. Consequently, the attrition rate decreases to 50%, compared with 75% for the full subset.

By applying this approach to the NPHS, the maximum potential sample after eight cycles of collection reaches 120,932 (7 x 17,276) analytical units (the NPHS contains 17,276 respondents multiplied by 7 analytical units possible after eight cycles). After eight cycles of collection, attrition reduces the number of potential analytical units to 95,046—to 79% of the original sample resulting in an attrition rate of 21%.

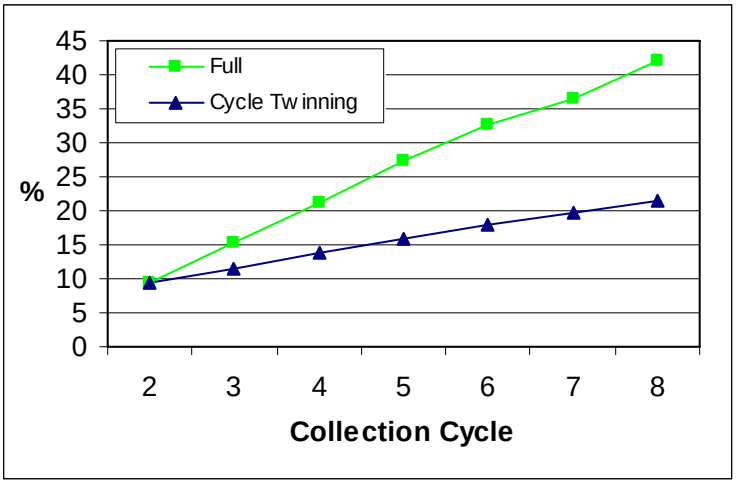
In comparison, the attrition rate for the full subset is 42% after eight cycles. By taking the analytical units, that result from using the cycle twinning approach, the attrition rate diminishes by approximately 21 percentage points in comparison to that observed with the

full subset (see Table 7.B and Figure 1).

Table 7.B: Attrition Rates

Cycle	Subset (%)	
	Full	Cycle twinning approach
2	9.3	9.3
3	15.3	11.4
4	21.3	13.7
5	27.3	15.9
6	32.7	18.0
7	36.4	19.6
8	42.2	21.4

Figure 1: NPHS Attrition Rates Taken from Table 7.B by Subset and by Collection Cycle



13.2 Creation of the Modified Subset

Before being able to analyze the data using the cycle twinning approach, the structure of the NPHS square subset must be modified. The NPHS Square subset contains 17,276 records or analytical units where each record represents information from all cycles for a single respondent. The transformation of this subset consists of modifying the analytical unit so that each unit consists of two consecutive cycles from a single respondent. For Cycle 8, the number of records for this modified subset would be 95,046 (see Section 7.1).

Using the example from Table 7.A, the record from respondent A is re-written to 7 analytical units, respondents B, C and D to 2, 4 and 1 analytical units, respectively (see Table 7.C). Here are the steps to follow for modification of the NPHS square subset and for conducting analysis using the cycle twinning approach.

Steps to follow for the* creation *of the NPHS modified subset:

Step 1: Calculate the possible number of twinnings for two consecutive cycles for each of the 17,276 respondents.

Step 2: Clone each NPHS respondent the same number of times as the number of twinnings calculated in step 1.

Step 3: Identify the longitudinal variables of interest for the twinned cycles.

Step 4: Identify the sample weights from Cycle 1: the longitudinal square weights (WT64LS), as well as the corresponding bootstrap weights (B5long) (see Sections 11.1 and 11.2).

Step 5: Keep only the information resulting from steps 3, 4 and 5. This step will increase the efficiency of analytical programs.

Table 7.C: Example of the Modified Subset Structure for Twinning

Respondent	Twinned cycles	Cycle 1 weight	Bootstrap weights	Variables of interest
A	1 and 2	W_A	$BS1_A$ à $BS500_A$	
A	2 and 3	W_A	$BS1_A$ à $BS500_A$	
A	3 and 4	W_A	$BS1_A$ à $BS500_A$	
A	4 and 5	W_A	$BS1_A$ à $BS500_A$	
A	5 and 6	W_A	$BS1_A$ à $BS500_A$	
A	6 and 7	W_A	$BS1_A$ à $BS500_A$	
A	7 and 8	W_A	$BS1_A$ à $BS500_A$	
B	1 and 2	W_B	$BS1_B$ à $BS500_B$	
B	5 and 6	W_B	$BS1_B$ à $BS500_B$	
C	1 and 2	W_C	$BS1_C$ à $BS500_C$	
C	2 and 3	W_C	$BS1_C$ à $BS500_C$	
C	3 and 4	W_C	$BS1_C$ à $BS500_C$	
C	4 and 5	W_C	$BS1_C$ à $BS500_C$	
D	3 and 4	W_D	$BS1_D$ à $BS500_D$	

13.3 Methodological Aspects of the Cycle Twinning Approach

As mentioned in Section 10.3, the sampling design and the sampling probabilities have an impact on the estimation method and variances calculations. The cycle twinning approach brings with it an additional layer of complexity: analytical units from the same respondent are highly correlated to each other.

With the help of ‘marginal models’ (Fitzmaurice et al. 2004), it is enough to consider the analytical unit resulting from the cycle twinning approach as an additional stage in the NPHS sampling design in to account for the dependency between the analytical units in the variance calculation. The correlation between the analytical units from the same respondent is therefore included in the sample design effect. Creating the modified subset

as described above and using the bootstrap method for variance calculation will ascertain the analyst to compute the variance correctly.

The choice of the variance calculation method is even more important given the high correlation between analytical units from the same respondent. The use of one of the methods described in Section 10.3, like the bootstrap method (SAS and SPSS) or other software such as STATA or SUDAAN is highly recommended.

13.4 An Example of How to Use the Cycle Twinning Approach: Quitting Smoking

Quitting smoking is one of the most important steps that smokers can take to improve their health. Understanding the factors that are associated with smoking cessation is important for public health programs aimed at reducing the smoking rate. Shields (2005)[25] used the NPHS to analyse factors associated with smoking cessation among people aged 18 or older who were daily smokers.

For this analysis, Cycles 1 to 5 of the NPHS were used. The NPHS follows the same sample of individuals over time, interviewing them at two-year intervals. An analysis file was assembled by examining successive pairs of NPHS cycles. The target population of interest was all respondents aged 18 or older who reported that they were daily smokers. Daily smokers were considered to be quitters if, in the following cycle, they reported not smoking at all in the last two years.

The analysis file was created by identifying daily smokers aged 18 or older at four baseline cycles (1, 2, 3, and 4) and determining if they had quit at the follow-up cycle two years later (i.e., Cycles 2, 3, 4, and 5, respectively). For a record to be written to the analyses file, two conditions had to be satisfied:

- the individual was a daily smoker aged 18 or older at the baseline cycle[26]
- smoking status was known at the follow-up cycle two years later (i.e., it was known if the individual still smoked or had quit).

The analysis file used contains many individuals who contributed more than one record, because they reported that they were daily smokers in more than one survey cycle.

To illustrate, an artificial example is shown below. Suppose that only two risk factors are studied to model the probability of quitting smoking: the number of cigarettes smoked per day (x) and the age of smoking initiation (z). Let y denote a binary variable that takes on the value 1 if a daily smoker had stopped smoking in the subsequent cycle two years later; otherwise y has the value 0. Information about x, y, and z is collected at each cycle.

Let p denote the probability that y takes the value 1; i.e., the individual has stopped smoking. Suppose that the model of interest is the logistic model

$$\log\left(\frac{p}{1-p}\right) = \alpha_0 + \alpha_1 x + \alpha_2 z\,,$$

25. Shields, M. (2005). The journey to quitting smoking. *Health Reports,* Vol. 16, No. 3. Statistics Canada, Catalogue 82-003.
26. In a pair of twinned cycles the baseline cycles refers to the first cycle of the pair. For example, for twinned cycles 3 and 4 the baseline cycle is Cycle 3.

and that the objective of the study is to estimate α_0, α_1 and α_2 and standard errors of their estimates. Suppose that only the first four cycles of the NPHS are being used for this analysis, and there are only four individuals on the NPHS file (all men) resulting in the following analysis file:

Cycle	Person ID	*x*	*z*	*y*[27]
1	1	x_{11}	z_{11}	0
1	2	x_{21}	z_{21}	1
1	3	x_{31}	z_{31}	1
1	4	x_{41}	z_{41}	0
2	1	x_{12}	z_{12}	0
2	4	x_{42}	z_{42}	0
3	1	x_{13}	z_{13}	0
3	3	x_{33}	z_{33}	1
3	4	x_{43}	z_{43}	1

- Person 1, who was a daily smoker in all 4 cycles, contributes 3 records to the analysis file since he was 18 or older and a daily smoker for all three baseline cycles (1, 2 and 3).
- Person 2 contributes only one record. He was a daily smoker at Cycle 1, had quit at Cycle 2, and did not respond to the survey at Cycles 3 and 4.
- Person 3 contributes two records. He was a daily smoker in Cycle 1, quit smoking in Cycle 2, took up smoking again in Cycle 3, and quit by Cycle 4.
- Person 4 contributes three records. He was a daily smoker in the first three cycles and then quit in Cycle 4.

This pattern of responses can be summarized:

Person ID	Cycle 1	Cycle 2	Cycle 3
1	X	X	X
2	X		
3	X		X
4	X	X	X

Suppose that the analysis is conducted using a software package for design-based logistic regression that requires an input data file that includes both the survey weight and the corresponding bootstrap weights. The survey and bootstrap weights from cycle 1 are used.

[27] This variable is derived from the subsequent cycle two years later.

The analysis file must be merged with the bootstrap weight for cycle 1, and should have the following structure:

Record	Person	Wt	X	Z	Y	bs1	bs2	---
1	1	w_1	x_{11}	z_{11}	0	$bs1_1$	$bs2_1$	---
2	1	w_1	x_{12}	z_{12}	0	$bs1_1$	$bs2_1$	---
3	1	w_1	x_{13}	z_{13}	0	$bs1_1$	$bs2_1$	---
4	2	w_2	x_{21}	z_{21}	1	$bs1_2$	$bs2_2$	---
5	3	w_3	x_{31}	z_{31}	1	$bs1_3$	$bs2_3$	---
6	3	w_3	x_{33}	z_{33}	1	$bs1_3$	$bs2_3$	---
7	4	w_4	x_{41}	z_{41}	0	$bs1_4$	$bs2_4$	---
8	4	w_4	x_{42}	z_{42}	0	$bs1_4$	$bs2_4$	---
9	4	w_4	x_{43}	z_{43}	1	$bs1_4$	$bs2_4$	---

Where w_i is the survey weight variable for person i and bsj_i is the j^{th} corresponding bootstrap weight for the same person. The weights and bootstrap weights are the same for multiple observations from the same person.

This approach uses the same design information (weights and bootstrap weights) and the same software that would be used if the model of interest were being fit to only one observation per individual. The approach simply requires that an individual's weight and bootstrap weights be assigned to multiple observations for that individual.

The question about which weight and bootstrap weights variables are the most appropriate can be debated, but the weight variables from the first cycle of the survey are often a good choice. This allows the analyst to maximize the sample size by making use of partial information when an individual has not responded to all NPHS cycles.

Although additional complexities are associated with partial response within individuals, these are not discussed in this document.

Appendix A: NPHS Household Component, Changes to the Questionnaire for Cycle 8 (2008/2009)

1. Introduction

This appendix describes the changes between the Cycle 7 (2006/2007) and the Cycle 8 (2008/2009) questionnaires. Some questions from Cycle 7 were removed in Cycle 8 and other questions, not in Cycle 7, were added in Cycle 8. The changes between the two cycles are described in detail in point 3. Globally, the main additions to Cycle 8 are in the stress module (6 focus questions on self-esteem were brought back from Cycle 4, and 13 focus questions on sense of coherence were brought back from Cycle 3). The questions removed at Cycle 8 are 12 food choice questions from the nutrition module.

2. Changes to Questionnaire Structure

Apart from the questions added and removed at Cycle 8, the order of the questionnaire remained the same as in Cycle 7.

3. Changes to Core Content

In the following description, external question names from the questionnaires are used. Some internal question names may have been changed to ensure consistency throughout the questionnaire owing to deletion or addition of questions. The variable names for the master, share, and public files are created using the variable naming convention.

Sections without modifications

- Household Record Variables
- General Health (GH)
- Sleep (SL)
- Height and Weight (HW)
  - Body Image
- Nutrition (NU)
  - Supplement use
  - Fruit and vegetable consumption
  - Soft drink consumption
  - Milk consumption
- Preventive Health (PH)
- Health Care Utilization (HC)
  - Home Care
- Restrictions of Activities (RA)
- Chronic conditions (CC)
  - Food or Digestive Allergies
  - Other Allergies
  - Asthma
  - Fibromyalgia
  - Arthritis or Rheumatism excluding Fibromyalgia
  - Back Problems
  - High Blood Pressure

  - Migraine Headaches
  - Chronic Bronchitis or Emphysema
  - Diabetes
  - Epilepsy
  - Heart Disease
  - Cancer
  - Intestinal or Stomach Ulcers
  - Effects of a stroke
  - Urinary Incontinence
  - Bowel Disorder
  - Alzheimer's Disease or other Dementia
  - Cataracts
  - Glaucoma
  - Thyroid Condition
  - Other Long-term Condition
- Health Status (HS)
  - Vision
  - Hearing
  - Speech
  - Getting Around
  - Hands and Fingers
  - Feelings
  - Memory
  - Thinking
  - Pain and Discomfort
- Physical Activities (PA)
- UV Exposure (UV) (TU)
- Repetitive Strain (RP)
- Injuries (IJ)
- Stress (ST)
  - Ongoing Problems
  - Childhood and Adult Stressors
  - Work Stress
  - Mastery
- Medication Use (DG)
- Smoking (SM)
- Alcohol (AL)
- Mental Health (MH)
- Social Support (SS)
- Language (SD)
- Education (ED)
- Labour Force (LF)
  - Job Attachment
  - Job Search—Last 4 Weeks
  - Past Job Attachment
  - Job Description
  - Absence/Hours
  - Other Job

  - Weeks Worked
  - Looking for Work
- Income (IN)
- Food insecurity (FI)
- Provincial Health Number and Administration (AM)
  - Provincial Health Number
  - Administration

Sections with modifications

Nutrition (NU)
- Food choice
  - 12 Focus questions (NU_B_1A, NU_B_1C to NU_B_1E, NU_B_2A to NU_B_2C, NU_B_3A to NU_B_3D and NU_2_3G) from this sub-section of Cycle 7 were dropped in Cycle 8.

Stress (ST)
- Self-esteem:
  - 6 Focus questions (ST_C_E1A to ST_C_E1F). Those questions were brought back from Cycle 4.
- Sense of Coherence:
  - 13 Focus questions (ST_C_H1 to ST_C_H13). Those questions were brought back from Cycle 3.

4. Buy-in Content

No buy-in content in Cycle 8.

Appendix B: NPHS Household Component, Examples of Variables from Previous Interviews Used as Additional Information in Cycle 8 (2008/2009)

Blood Pressure; Mammography; Pap Smear Test	In Cycle 1 and Cycle 2, the respondent was asked whether he/she ever had his or her blood pressure taken (or ever had a mammography etc.). In Cycle 3 the questions were repeated; however, the respondent was probed when said that he or she has not had the test done and in the previous cycle reported the contrary. In Cycles 4, 5, 6, 7, and 8, if the respondent had reported that he or she had had the test performed in a previous interview, only the question on the last time it was done was asked.
Restriction of Activities	Information on whether or not the respondent had a disability in a previous interview was used. If the status changed, an explanation of that change was probed.
Chronic Conditions	For each respondent, response to selected chronic conditions (asthma, fibromyalgia, arthritis, high blood pressure, migraine headaches, diabetes, epilepsy, stomach or intestinal ulcers, or the effects of a stroke) in a previous interview were used to help explain change. If it was a newly acquired condition, the date of onset for the condition was captured.
Smoking	If a daily smoker had reported the age at which he/she started smoking daily during the last interview, that response was used in Cycle 8. For the occasional smoker or non-smoker in Cycle 8 who had reported smoking daily (or having ever smoked daily) during the last interview, a flag about daily smoking was re-input. If smoking status changed, an explanation of that change was probed.
Socio-demographic Characteristics	For all respondents, a flag indicating that country of birth had been collected was used. Language first learned and still spoken was asked again because it can change over time.
Education	For all respondents, a flag indicating the highest level of education was re-input. Screening questions determined if the respondent was currently attending a learning institution between cycles. If so, educational attainment was collected anew.
Labour Force	For all respondents, the employer name, type of industry and duties of the main job from the previous interview were fed back. If the respondent indicated that they worked in the previous year, they were asked to confirm the employer name. When there was a change, the information was collected
Health Number	A flag indicated whether the health number that was collected in an earlier interview was valid. If the respondent's health number had not changed since last cycle or was invalid, the health number was requested again.